新世纪高等学校教材

环境科学与工程系列教材

HUANJING TONGJI FENXI

北京师范大学环境学院 组编

环境统计分析

杨晓华 刘瑞民 曾 勇 编著





新世纪高等学校教材

环境科学与工程系列教材

湿地学

环境水力学原理

环境影响评价实用教程

环境科学案例研究

环境科学案例研究教师手册

城市生态规划学

城市空气质量规划与管理

环境统计分析

环境与健康


ISBN 978-7-303-09502-5

9 787303 095025 >

定价：35.00元

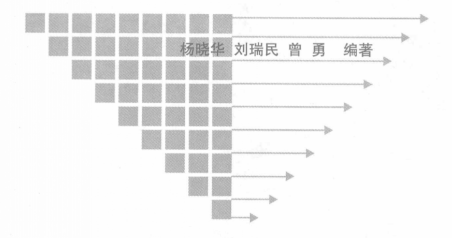
新世纪高等学校教材

环境科学与工程系列教材

北京师范大学环境学院 组编

环境统计分析

HUANJING TONGJI FENXI



杨晓华 刘瑞民 曾 勇 编著



图书在版编目(CIP)数据

环境统计分析/杨晓华,刘瑞民,曾勇编著.—北京:北京师范大学出版社,2008.8
(环境科学与工程系列教材)
ISBN 978-7-303-09502-5

Ⅰ.环… Ⅱ.①杨… ②刘… ③曾… Ⅲ.环境统计-统计分析-高等学校-教材 Ⅳ.X11

中国版本图书馆CIP数据核字(2008)第113056号

出版发行:北京师范大学出版社 www.bnup.com.cn
北京新街口外大街19号
邮政编码:100875
印　　刷:北京新丰印刷厂
经　　销:全国新华书店
开　　本:170mm×230 mm
印　　张:20.25
字　　数:337千字
印　　数:1~3 000册
版　　次:2008年9月第1版
印　　次:2008年9月第1次印刷
定　　价:35.00元

责任编辑:毛　佳　　装帧设计:高　霞
责任校对:李　菡　　责任印制:马鸿麟

前 言



环境统计分析是环境科学与环境工程的基础学科之一，是一门对环境系统不确定性问题进行数据处理、模型构建和分析的学科。环境系统，系指地球表面包括非生物、生物的各种环境因素及其相互关系的总和，是一个具有时、空、量、序变化的复杂巨系统。受人类活动、天文、气候和气象等众多因素的影响，环境系统中存在许多不确定性现象，并且有大量的数据需要进行统计分析和处理。环境的理论和实践对统计信息的需求急剧增加，对统计分析的理论和方法提出了更高的要求。在自然、社会与环境关系的基础上，用统计方法对环境问题予以量化描述和分析已成为环境研究的迫切需要。环境统计学的产生与发展使人们能够利用数理统计方法处理或解决环境中的不确定性问题，使其定量化，其中包括寻找变量之间的定量关系、从数据中发现环境趋势、探索环境系统变化规律。现代环境统计学一个很重要的标志就是模型技术的运用及量化分析。

全书分三大部分，共 10 章。其中，第 1 章属于基础篇，简要地介绍了环境统计分析的概率统计基础知识；第 2～9 章属于模型篇，阐述了环境一元线性回归分析、环境多元线性回归分析、环境系统聚类分析、环境模糊聚类分析、环境判别分析、环境主成分分析、环境因子分析、人工神经网络等方法、模型的原理，并给出了分析案例；第 10 章属于空间分析篇，介绍了环境空间统计分析的基本原理，

并给出了应用实例。全书的大多数例子都是用目前常用的统计分析语言 Matlab 编写实现的，是理论联系实际的经验总结，具有可操作性。本书适于做高等院校环境科学与环境工程专业的高年级本科生和研究生教材，对环境科学与环境工程、生态学、资源与管理、应用数学、地理科学等相关领域的学者和科研人员也有重要的参考价值。

本书第 1 章由杨晓华、曾勇执笔，第 2～9 章由杨晓华执笔，第 10 章由刘瑞民执笔，全书由杨晓华统稿。另外，尹心安参加了第 1 章的编写工作；王伟参加了第 3 章、第 4 章、第 10 章的编写工作；陈强、胡晓雪参加了第 5 章、第 6 章的编写工作；佘敦先参加了第 1 章、第 3 章、第 6 章、第 8 章的编写工作。2004 级、2005 级的博士研究生、2005 级的硕士研究生也提供了部分例题和习题。另外，习题答案均是用 Matlab 语言计算完成。

在本书的编写和出版过程中，北京师范大学环境学院院长杨志峰教授，副院长沈珍瑶、刘静玲教授，还有牛军峰、孙涛副教授以及北京师范大学出版社的胡廷兰、毛佳等同志对本书提出了许多宝贵意见。书中若干例题选自所列参考文献，在此一并表示感谢。由于我们的水平有限，书中错误在所难免，欢迎读者批评指正。

衷心感谢北京师范大学出版社给予的大力支持！

本书的完成得到国家重点基础研究发展规划项目（G2003CB415204）的资助，在此表示衷心的感谢！

编著者

2007 年 7 月

内 容 提 要

本书阐述了常用的环境统计分析方法，并给出了分析案例。首先简明扼要地介绍了环境统计分析的概率统计基础知识，又重点阐述了环境一元线性回归分析、环境多元线性回归分析、环境系统聚类分析、环境模糊聚类分析、环境判别分析、环境主成分分析和环境因子分析这些常用的环境统计分析模型；另外还给出了现代环境数据处理常用的人工神经网络方法和空间统计分析方法。对每一种方法，本书除了讲明基本原理外，还给出了大量的计算分析例题和案例。本书的部分例子是用目前实用的统计分析语言 Matlab 编写实现的，是理论联系实际的经验总结，具有实用性。本书适于做高等院校环境科学与环境工程专业的高年级本科生和研究生教材，对环境科学与环境工程、生态学、资源与管理、应用数学、地理科学等相关领域的学者和科研人员也有重要的参考价值。

目 录

部分思考题答案 ……………………………………… (270)

附录 …………………………………………………… (303)

第 1 章　概率统计基础

环境的理论和实践对统计信息的需求急剧增加，对统计分析的理论和方法提出了更高的要求。在自然、社会与环境关系的基础上，用统计方法对环境问题予以量化分析已成为环境科学工作者的迫切需要。环境统计学的产生与发展使人们能够利用数理统计方法处理或解决环境中的不确定性问题，使其定量化，其中包括寻找变量之间的定量关系、从数据中发现环境趋势、探索环境系统变化规律。为了能深刻理解和分析环境数据的数量特征和内在关系，需要我们首先掌握数理统计的基础知识。本章重点阐述环境统计分析的概率统计基础。

本章的主要内容是：

- 四种重要的概率分布；
- 随机向量的数字特征；
- 参数估计；
- 参数假设检验；
- 方差分析与试验设计初步。

1.1　四种重要的概率分布

在环境科学中，弄清统计分析对象的理论分布是关键的一环。土壤中的某些污染物、重金属的分布，大气中若干种微粒的浓度分布、监测值的误差分布等均服从正态分布或接近正态分布或取对数后服从正态分布。χ^2 分布、t 分布、F 分布是统计推断中经常碰到的另外三种分布。研究污染物在环境中的分布规律已是当前环境科学研究中重要的课题之一。

1.1.1　正态分布

市场上的食品很多是 1 kg 袋装，袋上标有“净含量 1 kg”的字样。但当用稍微精确一些的天平称那些食品的重量时，会发现有些可能会重些，有些可能会轻些，但都在 1 kg 左右。其中，多数离 1 kg 不远，离 1 kg 越近就越可能出现，离 1 kg 越远就越不可能。一般认为这种重量分布近似地服从正态分布(normal distribution)。近似地服从正态分布的变量很常见，如实验误差、商品的重量或

尺寸、某年龄人群的身高和体重等。在一定条件下，许多不是正态分布的样本均值在样本量很大时，也可用正态分布来近似。

若随机变量 X 的分布密度为：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty<x<+\infty,\ \sigma>0) \tag{1.1}$$

则称 X 服从正态分布 $N(\mu,\ \sigma^2)$，简记为 $X\sim N(\mu,\ \sigma^2)$。其中，μ 为均值，σ 为标准差，σ^2 为方差（标准差的平方）。

正态分布的密度曲线是一个对称的、呈钟形的曲线(最高点在均值处)(图 1-1)。正态分布是一族分布，各种正态分布根据它们的均值和标准差不同而有区别。标准差为 1 的正态分布 $N(0,\ 1)$称为标准正态分布(standard normal distribution)。标准正态分布的密度函数与分布函数记为：

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad (-\infty<x<+\infty) \tag{1.2}$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{y^2}{2}}dy \qquad (-\infty<x<+\infty) \tag{1.3}$$

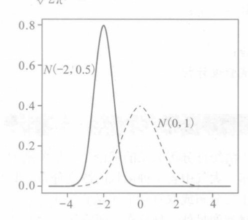


图 1-1　两条正态分布的密度曲线图
(左边是 $N(-2,\ 0.5)$分布，右边是 $N(0,\ 1)$分布)

在实际的生活中，我们经常会因为标准正态分布的优异特性而需要将一般的正态分布标准化，下面简单介绍一下正态分布的标准化过程。

设 $X\sim N(\mu,\ \sigma^2)$，作简单变换(减去其均值 μ，再除以标准差 σ)，则很容易得到随机变量 $Y=\frac{X-\mu}{\sigma}\sim N(0,\ 1)$。

因为：

$$E(Y)=E\left(\frac{X-\mu}{\sigma}\right)=\frac{1}{\sigma}[E(X)-\mu]=0$$

$$D(Y)=D\left(\frac{X-\mu}{\sigma}\right)=\frac{1}{\sigma^2}D(X)=1$$

这样就将一个普通的正态分布变成了一个标准的正态分布。

标准正态分布中还有一个十分重要的概念就是分位点。为了便于今后应用，对于标准正态随机变量，本书引入上侧分位点的定义(盛聚，1998)。

设 $X\sim N(0,1)$，若 z_α 满足条件

$$P(X>z_\alpha)=\alpha \qquad (0<\alpha<1)$$

则称 z_α 为标准正态分布的上侧 α 分位点，如图 1-2 所示。

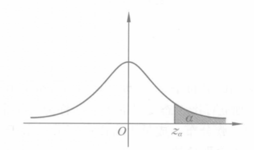


图 1-2 标准正态分布的上侧 α 分位点 z_α

例如，查附表 1 可知：$z_{0.01}=2.326\ 348$，$z_{0.05}=1.644\ 854$，$z_{0.10}=1.281\ 552$，$z_{0.154}=1.019\ 428$。

例 1.1 已知 $X\sim N(\mu,\sigma^2)$，求 X 在区间$(\mu-k\sigma,\ \mu+k\sigma)$的概率，这里 $k=1,2,3$。

解 $\forall a,b,0<a<b$，有：

$$P(a<X<b)=\int_a^b \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x$$

$$\xlongequal{t=\frac{x-\mu}{\sigma}}\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}\mathrm{d}t=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

这样在区间$(\mu-k\sigma,\ \mu+k\sigma)$的概率$(k=1,2,3)$为：

$$P(\mu-\sigma<X<\mu+\sigma)=\Phi(1)-\Phi(-1)=0.682\ 6$$
$$P(\mu-2\sigma<X<\mu+2\sigma)=\Phi(2)-\Phi(-2)=0.954\ 4$$
$$P(\mu-3\sigma<X<\mu+3\sigma)=\Phi(3)-\Phi(-3)=0.997\ 4$$

其中，$\Phi(-x)=1-\Phi(x)$。由此我们可以知道，属于正态分布的随机变量 X 之值，几乎都落在$(\mu-3\sigma,\ \mu+3\sigma)$区间里，落在该区间外的机会极少。

例 1.2 某地水体 COD 浓度 $X\sim N(5,2^2)$，求 COD 浓度落在区间(4，8)的

概率。

解 $\mu=5$，$\sigma=2$

$$\begin{aligned}P(4<X<8)&=\Phi\left(\frac{8-\mu}{\sigma}\right)-\Phi\left(\frac{4-\mu}{\sigma}\right)\\&=\Phi\left(\frac{8-5}{2}\right)-\Phi\left(\frac{4-5}{2}\right)\\&=\Phi(1.5)-\Phi(-0.5)\\&=0.9332-0.3085\\&=0.6247\end{aligned}$$

1.1.2 χ^2 分布

一个由正态变量导出的分布是 χ^2 分布(chi-square distribution)。该分布在一些检验中会用到。n 个独立标准正态变量的平方和称为有 n 个自由度的 χ^2 分布，记为 $\chi^2(n)$。χ^2 分布为一族分布，成员由自由度区分。由于 χ^2 分布变量为正态变量的平方和，因此它不会取负值。

设 X_1，X_2，…，X_n 是取自标准正态总体 $N(0,1)$的容量为 n 的样本，那么 $\chi^2=\sum_{i=1}^{n}X_i^2$ 即为由正态分布导出的自由度为 n 的 $\chi^2(n)$分布。所谓自由度，就是指可以自由取值的数据的个数，或者指不受任何约束、可以自由变动的变量的个数。

对于任意一个 $\chi^2(n)$分布，它的概率密度函数为：

$$P(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & (x>0)\\ 0 & (x\leqslant 0)\end{cases}\tag{1.4}$$

记为 $\chi^2\sim\chi^2(n)$，式中 n 为正整数，$\Gamma\left(\frac{n}{2}\right)$为 Γ 函数值，$\Gamma(z)=\int_0^{\infty}x^{z-1}e^{-x}dx$ $(z>0)$。

同正态分布类似，对于 χ^2 分布也有上侧 α 分位点。如果 $P(\chi^2>\chi^2_\alpha(n))=\alpha$，则称 $\chi^2_\alpha(n)$为上侧 α 分位点。对于不同的 α，n，上侧 α 分位点的值已制成表格（附表 3），可以查到。例如对于 $\alpha=0.050$，$n=9$，查得 $\chi^2_{0.050}(9)=16.919$。但大部分书只给出到 $n=45$ 的上侧 α 分位点的值。费歇尔(R. A. Fisher)曾证明，当 n 充分大时，近似有

$$\chi^2_\alpha(n)\approx\frac{1}{2}(z_\alpha+\sqrt{2n-1})^2\tag{1.5}$$

其中，z_α 为标准正态分布的上侧 α 分位点。利用式(1.5)可以求当 $n>45$ 时，$\chi_\alpha^2(n)$分布的上侧 α 分位点的近似值。

例如，查附表 1 并计算，可得：

$$\chi_{0.010}^2(100)\approx\frac{1}{2}(2.326\ 348+\sqrt{199})^2\approx135.023\ 1$$

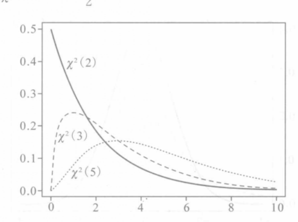


图 1-3　自由度分别为 2,3,5 的 χ^2 分布密度曲线图

1.1.3　*t* 分布

设 $X\sim N(0,\ 1)$，$Y\sim\chi^2(n)$，并且 X，Y 独立，则随机变量

$$t=\frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布(t-distribution 或 student's t)，记为 $t\sim t(n)$。

对于任意一个 $t(n)$分布，它的概率密度函数为：

$$P(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\qquad(-\infty<x<+\infty)\tag{1.6}$$

式中，n 为正整数。

不同的样本量通过标准化所产生的 t 分布也不同，这样就形成一族分布。t 分布的分布曲线关于 $x=0$ 对称，它的密度曲线看上去有些像标准正态分布，但是中间瘦一些，而且尾巴长一些。当自由度 k 无限增大时，t 分布将趋近于标准正态分布 $N(0,\ 1)$。

同样，类似于前面的两个分布，t 分布也有上侧 α 分位点的概念。

如果 $P(t>t_\alpha(n))=\alpha$，则称 $t_\alpha(n)$ 为 t 分布的上侧 α 分位点；$t_{1-\alpha}(n)=-t_\alpha(n)$，$t$ 分布的上侧 α 分位点，当 $n>45$ 时，可以用正态近似：

$$t_\alpha(n)\approx z_\alpha(n)$$

对于常用的 α 值，这样的近似值相对误差最多不超过 1.3%。

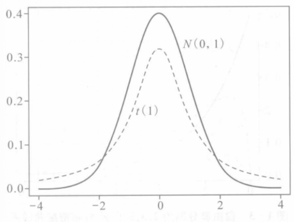


图 1-4 标准正态分布和 $t(1)$ 分布的密度曲线图

1.1.4 F 分布

设 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且 X，Y 相互独立，那么 $F=\dfrac{X/n_1}{Y/n_2}$ 称为自由度为(n_1, n_2)的 F 分布，n_1 和 n_2 分别称为第一自由度和第二自由度，通常记为 $F\sim F(n_1, n_2)$。

F 分布变量为两个 χ^2 分布变量（在除以它们各自自由度之后）的比；第一自由度等于在分子上的 χ^2 分布的自由度，第二自由度等于在分母上的 χ^2 分布的自由度。

对于任意一个 $F(n_1, n_2)$分布，它的概率密度函数为：

$$P(x)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}}x^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} & (x>0)\\ 0 & (x\leqslant 0)\end{cases} \tag{1.7}$$

F 分布的上侧 α 分位点定义为：如果 $P(F>F_\alpha(n_1, n_2))=\alpha$，则称 $F_\alpha(n_1, n_2)$

为 F 分布的上侧 α 分位点。实际上，对于 F 分布还有一个非常重要的性质，即：

$$F_{\alpha}(n_1, n_2)=\frac{1}{F_{1-\alpha}(n_2, n_1)} \tag{1.8}$$

一般情况下，F 分布表只给出了 $F_{\alpha}(n_1, n_2)$ 的值，而没有给出 $F_{1-\alpha}(n_2, n_1)$ 的值。此时，只要利用上述公式，就可以通过转换由 $F_{\alpha}(n_1, n_2)$ 求出 $F_{1-\alpha}(n_2, n_1)$。

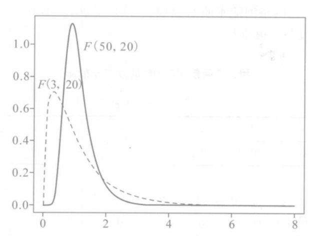


图 1-5 自由度为(3, 20)和(50, 20)的 F 分布密度曲线图

1.2 随机向量的数字特征

现代统计学以随机变量为研究对象，现代统计分析方法更涉及随机向量，下面先对概率论中随机向量的主要数字特征作一介绍。

1.2.1 数学期望

“平均数”是我们日常生活中使用最多的一个数字特征，如平均身高、平均浓度、平均产量、平均产值、平均成绩等。它简洁明了地指出所研究对象的位置特征，对评判事物、作出决策都有重要的作用。而数学期望实际上是以概率为权重的加权平均值。

1.2.1.1 离散随机变量的数学期望

在概率统计中，设 X 为离散型随机变量，它取得的一切可能值为 $x_k(k=1,$

2，…），对应的概率为 $P(X=x_k)=p_k(k=1, 2, \cdots)$，如果 $\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛，则称 $\sum_{k=1}^{\infty} x_k p_k$ 为随机变量 X 的数学期望，记作 $E(X)$，即：

$$E(X)=\sum_{k=1}^{\infty} x_k p_k \quad (k=1, 2, \cdots) \tag{1.9}$$

例 1.3 甲、乙两条河流水质监测项目 DO（溶解氧）的概率分布如表 1.1 所示，试问哪条河流污染更重？

表 1.1　　甲、乙两条河流 DO 的概率分布表

河流甲

X	0	1	2	3
P	0.3	0.3	0.2	0.2

河流乙

Y	0	1	2	3
P	0.3	0.5	0.2	0.0

解

$$E(X)=0\times0.3+1\times0.3+2\times0.2+3\times0.2=1.3$$
$$E(Y)=0\times0.3+1\times0.5+2\times0.2+3\times0.0=0.9$$

上面结果表明，河流乙污染更重。

1.2.1.2　连续随机变量的数学期望

设 X 是一个连续型的随机变量，密度函数为 $f(x)$，当 $\int_{-\infty}^{+\infty}|x|f(x)\mathrm{d}x<\infty$ 时，则称 X 的数学期望存在，且

$$E(X)=\int_{-\infty}^{+\infty} xf(x)\mathrm{d}x \tag{1.10}$$

例 1.4 设服从拉普拉斯分布的随机变量 X 的概率密度为

$$f(x)=\frac{1}{2}\mathrm{e}^{-|x|} \qquad (-\infty<x<+\infty)$$

求 $E(X)$。

解　$$E(X)=\int_{-\infty}^{+\infty} xf(x)\mathrm{d}x=\int_{-\infty}^{+\infty} x\,\frac{1}{2}\mathrm{e}^{-|x|}\,\mathrm{d}x$$

因为　奇函数在对称区间的积分为 0

所以　$E(X)=0$

1.2.1.3　随机变量函数的数学期望

设 X 为离散型随机变量，它取得的一切可能值为 $x_k(k=1, 2, \cdots)$，对应的

概率为 $P(X=x_k)=p_k$ （$k=1, 2, \cdots$），对于 X 的函数 $Y=g(X)$，如果 $\sum_{k=1}^{\infty}|g(x_k)|p_k$ 收敛，则 Y 的数学期望为：

$$E(Y)=E[g(X)]=\sum_{k=1}^{\infty}g(x_k)p_k \tag{1.11}$$

同理，设 X 是一个连续的随机变量，密度函数为 $f(x)$，则对于任意一个关于 X 的函数 $Y=g(X)$，则 Y 的数学期望定义为：

$$E(Y)=E[g(X)]=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x \tag{1.12}$$

例 1.5　设随机变量 X 的概率分布如表 1.2 所示，且 $Y_1=2X+1$，$Y_2=X^2$，求 $E(Y_1)$和 $E(Y_2)$。

表 1.2　　**随机变量 X 的概率分布表**

X	-2	-1	0	1
P_k	0.1	0.3	0.4	0.2

解　方法一：先求 Y_1 和 Y_2 的概率分布（表 1.3），再求 $E(Y_1)$和 $E(Y_2)$。

表 1.3　　**随机变量 Y_1 和 Y_2 的概率分布表**

Y_1	-3	-1	1	3
P	0.1	0.3	0.4	0.2

Y_2	0	1	4
P	0.4	0.5	0.1

由公式(1.9)有：

$$E(Y_1)=(-3)\times0.1+(-1)\times0.3+1\times0.4+3\times0.2=0.4$$

$$E(Y_2)=0\times0.4+1\times0.5+4\times0.1=0.9$$

方法二：直接由公式(1.11)求 $E(Y_1)$，$E(Y_2)$。

$$E(Y_1)=E(2X+1)=[2\times(-2)+1]\times0.1+[2\times(-1)+1]\times0.3+[2\times0+1]\times0.4+[2\times1+1]\times0.2=0.4$$

$$E(Y_2)=E(X^2)=(-2)^2\times0.1+(-1)^2\times0.3+0^2\times0.4+1^2\times0.2=0.9$$

1.2.1.4　性质

数学期望具有以下几个性质(a，b，c 均为常数)：

(1)$E(c)=c$；

(2)$E(aX)=aE(X)$；

(3)$E(aX+b)=aE(X)+b$。

1.2.2 方差和均方差

1.2.2.1 方差和均方差的定义

数学期望描述随机变量可能值的集中位置，在实用上还需要了解随机变量可能值离散程度，这就需要引入一个新的特征数——方差。我们称随机变量 X 与其平均值差的平方的期望值称为随机变量 X 的方差，记作 $D(X)$，即：

$$D(X)=E[X-E(X)]^2 \tag{1.13}$$

随机变量方差的平方根称为均方差或标准差，记作 $\sigma=\sqrt{D(X)}$。

由方差的定义可知，方差是一个非负数，方差的大小刻画了随机变量取值的分散程度。

1.2.2.2 离散型随机变量的方差

设 X 为离散型随机变量，它的概率分布为 $P(X=x_k)=p_k(k=1, 2, \cdots)$，则 X 的方差表达式为：

$$D(X)=\sum_{k=1}^{\infty}[x_k-E(X)]^2 p_k \tag{1.14}$$

例 1.6 计算例 1.3 中的河流水质指数溶解氧的方差。

解 由于：

$$\begin{aligned}D(X)&=E[X-E(X)]^2\\&=E\{X^2-2XE(X)+[E(X)]^2\}\\&=E(X^2)-2E(X)E(X)+[E(X)]^2\\&=E(X^2)-[E(X)]^2\end{aligned}$$

因此：

$$E(X^2)=0^2\times0.3+1^2\times0.3+2^2\times0.2+3^2\times0.2=2.9$$

$$D(X)=E(X^2)-[E(X)]^2=2.9-1.3^2=1.21$$

$$E(Y^2)=0^2\times0.3+1^2\times0.5+2^2\times0.2+3^2\times0.0=1.3$$

$$D(Y)=E(Y^2)-[E(Y)]^2=1.3-0.9^2=0.49$$

1.2.2.3 连续型随机变量的方差

设 X 为连续型随机变量，它的概率密度函数为 $P(x)$，则 X 的方差表达式为：

$$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2P(x)\mathrm{d}x \tag{1.15}$$

例1.7　X 服从参数为 λ 的指数分布，即 X 的概率密度函数为：

$$P(x)=\begin{cases}\lambda \mathrm{e}^{-\lambda x} & (x\geqslant 0)\\ 0 & (x<0)\end{cases}$$

求 $D(X)$。

解

$$E(X^2)=\int_{-\infty}^{+\infty}x^2P(x)\mathrm{d}x=\int_{0}^{+\infty}\lambda x^2\mathrm{e}^{-\lambda x}\mathrm{d}x$$

$$=\left(-x^2\mathrm{e}^{-\lambda x}-\frac{2x}{\lambda}\mathrm{e}^{-\lambda x}-\frac{2}{\lambda^2}\mathrm{e}^{-\lambda x}\right)\Bigg|_0^{+\infty}=\frac{2}{\lambda^2}$$

$$D(X)=E(X^2)-[E(X)]^2=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2}$$

1.2.2.4　性质

方差具有以下几个性质(a，b，c 均为常数)：

(1)$D(c)=0$；

(2)$D(aX)=a^2D(X)$；

(3)$D(aX+b)=a^2D(X)$；

(4)$D(X)=0$ 的充要条件是存在常数 c，使得 $P\{X=c\}=1$。

此外，方差还有一个重要的等式：$D(X)=E(X^2)-[E(X)]^2$。

1.2.3　原点矩和中心矩

除了数学期望和方差外，在研究随机变量时还经常用到随机变量的各阶矩——原点矩和中心矩。例如在定义了随机变量的数学期望、方差这些数字特征之后，如果记数学期望为 $E(X)$，则方差为：

$$D(X)=E(X^2)-[E(X)]^2$$

对于离散型随机变量和连续型随机变量来说，分别有如下的公式：

$$E(X)=\sum_{i=1}^{\infty}x_ip_i;\ E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x \tag{1.16}$$

$$E(X^2)=\sum_{i=1}^{\infty}x_i^2p_i;\ E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)\mathrm{d}x \tag{1.17}$$

这些计算式与物理学中静力矩和惯性力的计算式相似，借用物理学中“矩”的名字，$E(X)$和 $E(X^2)$分别称为一阶矩和二阶矩。对任意正整数 k，可以自然

地定义 $E(X^k)$ 为随机变量 X 的 k 阶矩。

注意到方差 $D(X)=E[X-E(X)]^2$，当然也可以对任意的正整数 k，考虑

$$E\ [X-E(X)]^k \tag{1.18}$$

它也是一种 k 阶矩。实际上，$E(X)$ 是 X 的一个中心，因而常常把 $E[X-E(X)]^k$ 称为随机变量 X 的 k 阶中心矩。而 $E(X^k)$ 是对原点的 k 阶矩，也就称之为 k 阶原点矩。

1.2.4 变异系数

如果两组数据的计量单位相同，并且均值一样，可以利用标准差来比较两组数据的离散程度。但当两组数据的计量单位不同或者均值不同时，就不能直接比较两组数据的标准差来分析两组数据的离散程度。由此引入变异系数 C_v，它的定义如下：

设随机变量 X 的标准差为 σ，数学期望为 $E(X)$，则标准差与数学期望的比值称为变异系数，记为：

$$C_v=\frac{\sigma}{E(X)} \tag{1.19}$$

例如下面两组数据(4，5，6，7，8)与(40，50，60，70，80)的标准差分别是 1.58 和 15.8，如果仅从标准差来看显然第二组数据分散程度较大。但是由于两组数据的均值不同，分别为 6 和 60，单纯由标准差来判断数据的分散程度就不合适。实际上，上述两组数据的变异系数，均为：$C_v=0.26$。因此，两组数据的分散程度是相同的。

1.2.5 协方差阵和自协方差阵

对于二维随机向量(X，Y)，它的一个重要的数字特征是协方差，若 X 与 Y 的 1+1 阶混合中心矩存在，记为 $\mathrm{Cov}(X,Y)$，即：

$$\mathrm{Cov}(X,Y)=E\{[X-E(X)][Y-E(Y)]\} \tag{1.20}$$

这里，如果随机变量 X 和随机变量 Y 为同一个变量，则

$$\mathrm{Cov}(X,X)=E\{[X-E(X)][X-E(X)]\}=E[X-E(X)]^2=D(X)$$

设 $X=(X_1, X_2, \cdots, X_n)$ 和 $Y=(Y_1, Y_2, \cdots, Y_p)$ 分别为 n 维和 p 维的随机向量，它们之间的协方差阵定义为一个 $n\times p$ 矩阵，其元素为 $\mathrm{Cov}(X_i, Y_j)$ ($i=1, 2, \cdots, n$; $j=1, 2, \cdots, p$)，即：

$$\mathrm{Cov}(X,Y)=\begin{pmatrix}\mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \cdots & \mathrm{Cov}(X_1,Y_p)\\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \cdots & \mathrm{Cov}(X_2,Y_p)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_n,Y_1) & \mathrm{Cov}(X_n,Y_2) & \cdots & \mathrm{Cov}(X_n,Y_p)\end{pmatrix} \tag{1.21}$$

如果 $\mathrm{Cov}(X,Y)=0$，则称 X 和 Y 是不相关的。

同样，如果随机向量 X 和随机向量 Y 是同一个随机向量，那么 $\mathrm{Cov}(X,Y)$ 就相应转化为如下形式：

$$\mathrm{Cov}(X,X)=\begin{pmatrix}D(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n)\\ \mathrm{Cov}(X_2,X_1) & D(X_2) & \cdots & \mathrm{Cov}(X_2,X_n)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & D(X_n)\end{pmatrix} \tag{1.22}$$

称上述矩阵为随机向量 X 的自协方差阵。

1.2.6 随机变量的相关系数

设随机变量 X 和 Y 的期望与方差都存在，且 $D(X)>0$，$D(Y)>0$，把相关系数 ρ_{XY} 定义为：

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{1.23}$$

相关系数描述了随机变量之间的相关程度。

对于给定的任意两个随机变量 X 和 Y，它们之间的相关系数 ρ_{XY}，总是满足 $|\rho_{XY}|\leqslant 1$，并且 $|\rho_{XY}|=1$ 的充要条件是 X 和 Y 线性相关，即存在常数 a 和 b，使得 $Y=aX+b$。

相关系数只是随机变量间线性关系强弱的一个度量，因而说得更确切些，应该把它叫做线性相关系数。

例 1.8 考察某河流的有机污染状况，分别在 4 个不同断面上监测了 BOD_5、COD、DO 三项指标（表 1.4），试求协方差矩阵和相关系数矩阵。

表 1.4 某河流水质监测结果 单位：mg/L

断面	BOD_5	COD	DO
1#	2	3	8
2#	3	5	8
3#	19	19	4
4#	3	6	7

解 根据协方差定义，计算得表1.5。

表1.5 协方差阵元素所需数据表

m	X_i	X_j	X_k	$X_i-\overline{X}_i$	$X_j-\overline{X}_j$	$X_k-\overline{X}$
1	2	3	8	−4.75	−5.25	1.25
2	3	5	8	−3.75	−3.25	1.25
3	19	19	4	12.25	10.75	−2.75
4	3	6	7	−3.75	−2.25	0.25
$\overline{X}$	6.75	8.25	6.75			

$$\sigma_{X_iX_j}=E\{[X_i-E(X_i)][X_j-E(X_j)]\}=\frac{1}{n}\sum_{m=1}^{n}(X_{mi}-\overline{X}_i)(X_{mj}-\overline{X}_j)$$
$$=44.312\ 5$$

$$\sigma_{X_iX_k}=E\{[X_i-E(X_i)][X_k-E(X_k)]\}=\frac{1}{n}\sum_{m=1}^{n}(X_{mi}-\overline{X}_i)(X_{mk}-\overline{X}_k)$$
$$=-11.312\ 5$$

$$\sigma_{X_jX_k}=E\{[X_j-E(X_j)][X_k-E(X_k)]\}=\frac{1}{n}\sum_{m=1}^{n}(X_{mj}-\overline{X}_j)(X_{mk}-\overline{X}_k)$$
$$=-10.187\ 5$$

$$\sigma_{X_iX_i}=E\{[X_i-E(X_i)]^2\}=\frac{1}{n}\sum_{m=1}^{n}(X_{mi}-\overline{X}_i)^2=50.187\ 5$$

$$\sigma_{X_jX_j}=E\{[X_j-E(X_j)]^2\}=\frac{1}{n}\sum_{m=1}^{n}(X_{mj}-\overline{X}_j)^2=39.687\ 5$$

$$\sigma_{X_kX_k}=E\{[X_k-E(X_k)]^2\}=\frac{1}{n}\sum_{m=1}^{n}(X_{mk}-\overline{X}_k)^2=2.687\ 5$$

即协方差阵为：

$$\boldsymbol{V}=\begin{pmatrix}50.187\ 5 & 44.312\ 5 & -11.312\ 5\\ 44.312\ 5 & 39.687\ 5 & -10.187\ 5\\ -11.312\ 5 & -10.187\ 5 & 2.687\ 5\end{pmatrix}$$

根据相关系数的定义，得相关系数矩阵为：

$$\boldsymbol{R}=\begin{pmatrix}1.000 & 0.993 & -0.974\\ 0.993 & 1.000 & -0.986\\ -0.974 & -0.986 & 1.000\end{pmatrix}$$

1.2.7 总体与样本

环境统计学研究的对象是环境现象的数量关系和数量特征，是关于数据收集、整理、归纳和分析的方法论科学，是环境研究的一种重要方法。环境统计方法广泛地运用于环境科学的各个方面。环境部门要作出决策、执行计划、检查监督、控制环境污染等都需要以充分、可靠的统计资料为基础。环境统计方法，需要在搜集环境数据的基础上，对数据整理和统计描述，运用统计方法及环境对象的有关知识，从定量与定性的结合上进行统计分析。环境统计描述可以把数据、情况、问题、建议等融为一体，是发挥环境统计的信息、咨询、管理、监督和决策功能的重要内容。

在统计分析中，我们把研究对象的全体所构成的集合称为总体（或母体），总体通常用 X，Y，Z 来表示。把组成总体的每一个成员称为个体。一个总体中所含的个体的数量称为总体的容量。在实际中，为了研究总体中个体的各种数值指标和推断总体的某些特征，总是通过对总体中部分个体的观测和实验来推测估计整个总体的特性，这样就需要从总体中按一定的抽样技术抽取若干个个体，通常将这一抽取过程称为抽样。所抽取的部分个体称为样本，样本中所含个体的数量称为样本容量。

通过一定的抽样技术从总体中抽取了一定的样本后，这些样本所含的信息并不能直接得到总体的特征，这样就需要我们对这些样本进行数学处理，构造不同的函数来反映总体的信息。在数理统计中，这样的函数被称为统计量。

统计量的定义：设总体构成的集合为 X，其中，X_1，X_2，…，X_n 为从总体中抽取的样本，$\varphi(X_1, X_2, \cdots, X_n)$是$(X_1, X_2, \cdots, X_n)$的一个函数，且 φ 中不含任何未知参数，则称 $\varphi(X_1, X_2, \cdots, X_n)$是一个统计量。例如，设$(X_1, X_2)$是从总体 $N(\mu, \sigma^2)$中抽取的一个二维样本，其中 σ 为未知参数，则式子 X_1-X_2，$X_1^2+X_2^2-3$，$X_1+2\mu X_2$ 为样本统计量；而$\frac{X_1}{\sigma}$，$\frac{1}{2}(X_1+X_2)-\sigma$ 则由于包含未知参数 σ，不是样本统计量。

在有些情况下，人们获得的统计资料并非事物整体的状况，而是来自事物的一个局部。如何利用局部的数据去推断整体的情况，以及这种推断的有效性和可靠性如何，即是推断统计所要研究的内容。通俗地说：用样本统计量去估计总体参数的依据是什么？要回答这个问题，就需要掌握统计学的重要定理：大数定理和中心极限定理。

1.2.8 样本子样的一些数字特征

设样本总体构成的集合为 X，其中，X_1，X_2，…，X_n 为从总体中抽取的一个子样，x_1，x_2，…，x_n 分别为相应的观测值，则：

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$$

仍称为样本均值；

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

仍称为样本方差，方差的平方根 s 称为样本的标准差；统计量

$$a^k=\frac{1}{n}\sum_{i=1}^{n}x_i^k \qquad (k=1,\ 2,\ \cdots)$$

称为样本的 k 阶矩(或 k 阶原点矩)；统计量

$$b^k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k \qquad (k=2,\ 3,\ \cdots)$$

称为样本的 k 阶中心矩。

1.2.9 大数定律

根据概率的定义，在相同条件下，重复进行 n 次试验，若在 n 次试验中，事件 A 发生的次数(可称为频数)为 n_A，则称比值 $f=\frac{n_A}{n}$ 为事件 A 在 n 次试验中发生的频率 f。如果当 n 充分大时，A 发生的频率稳定地在某一数值 p 附近摆动，而且一般来说随着试验次数的增多，这种摆动的幅度越变越小，即频率越来越稳定于 p，则称 p 为此随机试验中随机事件 A 发生的概率，记作：

$$P(A)=p$$

下面的大数定律将从理论上进一步证实事件的频率具有稳定性。

定理(伯努利大数定律) 设 n_A 是 n 次独立重复试验中事件 A 发生的次数，$p(0<p<1)$是在每次试验中事件 A 发生的概率，则对于任意正数 ε，有：

$$\lim_{n\to\infty}P\left\{\left|\frac{n_A}{n}-p\right|\geqslant\varepsilon\right\}=0$$

证明从略。显然，由互逆事件间的概率关系，上式又可写成如下的形式：

$$\lim_{n\to\infty}P\left\{\left|\frac{n_A}{n}-p\right|<\varepsilon\right\}=1$$

伯努利大数定律从理论上证明了事件 A 在 n 次独立重复试验中发生的频率 $\frac{n_A}{n}$，当 n 逐渐增大时稳定于事件 A 的概率 p。于是，当 n 充分大时，频率可以作为概率的近似值。

利用伯努利大数定理，可以近似求解随机变量的概率分布函数，下面介绍用直方图来表示频率直方图和累计频率图。

例 1.9　噪声测量结果(单位：dB)：50，51，52，52，53，53，53，54，54，55，求概率分布和概率密度曲线。

解　(1)找出数据中的最小值 m=50，最大值 M=55，极差为 M−m=5。

(2)数据分组，取 $a=49.5$(略小于 m)，$b=55.5$(略大于 M)，则所有样本值全部落入区间(a, b)内，将该区间分为 $k=6$ 等份，称每一等份的长度 $h=\frac{b-a}{k}=1$ 为组距；决定分组点，分组如下：49.5～50.5，50.5～51.5，51.5～52.5，52.5～53.5，53.5～54.5，54.5～55.5。

(3)作出频数、频率分布表（表 1.6）。

表 1.6　　**频数、频率分布表**

组序	区间范围	频数(f_i)	频率($W_j=f_j/n$)	累计频率(F_j)
1	49.5～50.5	1	0.1	0.1
2	50.5～51.5	1	0.1	0.2
3	51.5～52.5	2	0.2	0.4
4	52.5～53.5	3	0.3	0.7
5	53.5～54.5	2	0.2	0.9
6	54.5～55.5	1	0.1	1.0

(4)作出频率直方图和累计频率图。

以样本值为横坐标，频率(频率/组距，对于异距频率分布)为纵坐标的直角坐标系中，以分组区间为底，以 $Y_j=\frac{W_j}{x_{j+1}-x_j}=W_j$ ($j=1, 2, \cdots, 6$) 为高作一系列矩形，即频率直方图（图 1-6）。累计频率图，见图 1-7。

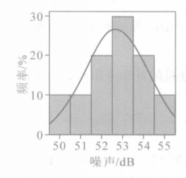


图 1-6 频率直方图

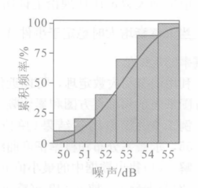


图 1-7 累计频率图

(5)通过矩形顶边画一条光滑的曲线，得到概率密度曲线和分布函数曲线的近似曲线（图 1-6～1-7）。

1.2.10 中心极限定理

假如从总体中随机抽取若干个容量为 n 的样本，对每个样本都可以计算均值，这些均值的分布即为样本均值的分布。下面我们通过一个算例引出一个重要的定理——中心极限定理(康永尚等，2005)。

例 1.10 设某一随机变量 X(总体)包含 5 个个体(样本点)，其变量值、期望值和方差列于表 1.7 中，随机抽取容量为 2 的样本，考察样本均值的分布。

表 1.7 随机变量 X(总体)的取值、期望值和方差

序号	X 的取值	X 的期望值	X 的方差
1	332		
2	336	$\mu=E(X)=\frac{1}{5}\sum_{i=1}^{5}x_i=340$	$D(X)=\sigma^2=\frac{1}{5}\sum_{i=1}^{5}(x_i-340)^2=32$
3	340		
4	344		
5	348		

解 随机抽取容量为 2 的样本(放回抽样)，可能抽到的样本点组合、变量值组合、均值及均值统计量列于表 1.8 中。

表 1.8　样本点组合、变量值组合、均值及均值统计量

样本序号	样本点组合	变量值组合	样本均值	样本均值统计量
1	(1，1)	(332，332)	332	样本均值的期望值：
2	(1，2)	(332，336)	334	
3	(1，3)	(332，340)	336	$E(\overline{X})=\frac{1}{25}\sum_{i=1}^{25}\bar{x}_i=340$
4	(1，4)	(332，344)	338	
5	(1，5)	(332，348)	340	样本均值的方差：
6	(2，1)	(336，332)	334	
7	(2，2)	(336，336)	336	$D(\overline{X})=\frac{1}{25}\sum_{i=1}^{25}[\bar{x}_i-E(\overline{X})]^2=16$
8	(2，3)	(336，340)	338	
9	(2，4)	(336，344)	340	
10	(2，5)	(336，348)	342	
11	(3，1)	(340，332)	336	
12	(3，2)	(340，336)	338	
13	(3，3)	(340，340)	340	
14	(3，4)	(340，344)	342	
15	(3，5)	(340，348)	344	
16	(4，1)	(344，332)	338	
17	(4，2)	(344，336)	340	
18	(4，3)	(344，340)	342	
19	(4，4)	(344，344)	344	
20	(4，5)	(344，348)	346	
21	(5，1)	(348，332)	340	
22	(5，2)	(348，336)	342	
23	(5，3)	(348，340)	344	
24	(5，4)	(348，344)	346	
25	(5，5)	(348，348)	348	

对比表 1.7 和表 1.8 可以发现，样本均值的期望值正好等于变量(总体)的期望值，样本均值的方差正好等于变量(总体)的方差除以样本容量。这些现象的出现不是偶然的，而是背后的定理发挥作用的结果，这一定理就是中心极限定理。

中心极限定理可以表述为：从期望值为 μ、方差为 σ^2 的随机变量(总体)中独立随机地抽取容量为 n 的样本，该样本的均值 $\overline{X}$ 为一随机变量，该随机变量服从于期望值为 μ、方差为$\frac{\sigma^2}{n}$的正态分布，即：

$$E(\overline{X})=\mu$$
$$D(\overline{X})=\frac{\sigma^2}{n} \tag{1.24}$$

且这种趋势随 n 的增大而愈加明显。

根据具体的应用条件，可分为三种情况：

(1)总体服从正态分布，且 σ^2 已知，此时，$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$服从标准正态分布。

(2)总体服从任意分布，σ^2 未知，$n\geqslant 30$，此时，$\frac{\overline{X}-\mu}{s/\sqrt{n}}$近似服从标准正态分布，其中 $s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n-1}}$ 为样本均方差。

(3)总体服从正态分布，σ^2 未知，$n<30$，此时，中心极限定理已不适合，$\frac{\overline{X}-\mu}{s/\sqrt{n}}$服从自由度为$(n-1)$的 Student 分布(又称 t 分布)，其中 $s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n-1}}$ 为样本均方差。

1.3 参数估计

估计就是根据拥有的信息来对现实世界进行某种判断。数理统计的基本任务是依据取自总体的样本对总体进行推断。如果知道了总体的分布类型，分布便由几个与总体有关的未知数字决定。要掌握总体的分布，依据样本对这些未知参数作尽可能准确的推断就显得非常重要。

从数据得到关于现实世界结论的过程就叫做统计推断(statistical inference)。估计(estimation)是统计推断的重要内容之一。统计推断的另一个主要内容是下一节要引进的假设检验(hypothesis test)。

人们往往先假定某数据来自一个特定的总体族(比如正态分布族)，但若要确定是总体族的哪个成员则需要知道总体参数值(比如总体均值和总体方差)，于是可以用相应的样本统计量(比如样本均值和样本方差)来估计相应的总体参数。

一些常见的涉及总体的参数包括总体均值(μ)、总体标准差(σ)或方差(σ^2)。正态分布族中的成员由(总体)均值和标准差完全确定。

估计的根据为从总体抽取的样本。如果样本已经得到，把数据代入之后，估计量就有了一个数值，称为该估计量的一个实现(realization)或取值，也称为一个估计值。

这里简单地介绍两种估计，一种是点估计(point estimation)，即用估计量的实现值来近似相应的总体参数；另一种是区间估计(interval estimation)，它是包括估计量在内(有时是以估计量为中心)的一个区间，该区间被认为很可能包含总体参数。点估计给出一个数字，用起来很方便；而区间估计给出一个区间。

最常用的估计量就是我们熟悉的样本均值、样本标准差；人们用它们来分别估计总体均值(μ)、总体标准差(σ)。

本书中只对参数估计部分作简单的介绍。

1.3.1　点估计

点估计是对真值θ以单一的数据$\hat{\theta}$为估计值的方法。用单一的数据表示估计值，在环境问题的分析及预测中经常用到。点估计问题就是要根据样本X_1，X_2，…，X_n构造一个统计量$\hat{\theta}(X_1, X_2, \cdots, X_n)$作为参数$\theta$的估计，我们称$\hat{\theta}$为参数$\theta$的估计量。如果$x_1$，$x_2$，…，$x_n$是样本的一组观测值，代入统计量$\hat{\theta}$就得到$\hat{\theta}$的具体数值，这个数值常称为$\theta$的估计值。估计量$\hat{\theta}$是样本$X_1$，$X_2$，…，$X_n$的函数，它不包含未知参数，也就是说$\hat{\theta}$是一个估计用的统计量。当我们获得样本观测值$x_1$，$x_2$，…，$x_n$后，就用$\hat{\theta}(x_1, x_2, \cdots, x_n)$作为未知参数$\theta$的估计值。对于不同的样本观测值，所得的估计值是不同的。例如：假定我们要估计一个班学生考试成绩的平均分数，根据一个抽出的随机样本计算的平均分数为80分，我们就用80分作为全班考试成绩平均分数的一个估计值，这就是点估计。

1.3.2　区间估计

因为点估计难以评价待估参数估计值与其真值之间的接近程度，即无法通过点估计来度量估计值的可信程度，因此引进区间估计。

给出一个区间(置信区间)($\hat{\theta}_1$，$\hat{\theta}_2$)，并预测真正的参数以一定的概率属于该区间的方法称为区间估计，这一区间能够覆盖真值的概率称为置信系数。当给定常数α($0<\alpha<1$)，若有$P(\hat{\theta}_1<\theta<\hat{\theta}_2)\geqslant 1-\alpha$成立，则称$\hat{\theta}_1$到$\hat{\theta}_2$这一区间能够覆

盖真值的概率为 $1-\alpha$。$(\hat{\theta}_1, \hat{\theta}_2)$为待估参数 θ 的置信水平为 $1-\alpha$ 的置信区间；$\hat{\theta}_1$，$\hat{\theta}_2$ 称为置信下限和置信上限；$1-\alpha$ 称为置信水平；α 称为显著性水平，为区间$(\hat{\theta}_1, \hat{\theta}_2)$不含 θ 的概率，即对未知参数估计失准的概率。

以下以正态分布样本均值的区间估计为例说明区间估计的含义。

根据中心极限定理，$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$服从标准正态分布，从而可以得到：

$$P\left\{-x_{\alpha/2}\leqslant\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\leqslant x_{\alpha/2}\right\}=1-\alpha \tag{1.25}$$

上式经过变换后可得：

$$P\left\{\overline{X}-\frac{\sigma}{\sqrt{n}}x_{\alpha/2}\leqslant\mu\leqslant\overline{X}+\frac{\sigma}{\sqrt{n}}x_{\alpha/2}\right\}=1-\alpha \tag{1.26}$$

上式即表示：在$(1-\alpha)$置信水平下总体期望值的区间估计为：

$$\overline{X}-\frac{\sigma}{\sqrt{n}}x_{\alpha/2}\leqslant\mu\leqslant\overline{X}+\frac{\sigma}{\sqrt{n}}x_{\alpha/2} \tag{1.27}$$

例 1.11 河流某一河段溶解氧(DO)含量符合 $X\sim N(\mu, \sigma^2)$，$\sigma^2=1$，今从该河段中随机监测了 5 次，监测结果(单位：mg/L)为 4，4，5，6，6，试在置信度 0.95 下，求参数 μ 的区间估计。

解 根据中心极限定理：$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$服从标准正态分布，故该河段溶解氧的期望值的置信区间可用公式 $\overline{X}-x_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\leqslant\mu\leqslant\overline{X}+x_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$求取。

因为：

$$\overline{X}=5,\ n=5,\ x_{0.025}=1.96$$

所以经计算，参数 μ 的置信度为 0.95 的区间估计为(4.123 5，5.876 5)。

统计学家想出了许多标准来衡量一个估计量的好坏。每个标准一般都仅反映估计量的某个方面。对一个估计量的衡量标准主要有无偏性、一致估计性和有效性。下面就简单介绍一下估计量的几个评价标准。

1.3.2.1 无偏估计量

设 $\hat{\theta}=\hat{\theta}(X_1, X_2, \cdots, X_n)$是未知参数 θ 的一个估计量，若

$$E(\hat{\theta})=\theta$$

则称 $\hat{\theta}$ 为 θ 的无偏估计量。

事实上对于无偏估计量，提出了 $\hat{\theta}$ 应该满足的无系统偏差的条件。$\hat{\theta}$ 是一个随机变量，其取值应在参数真值 θ 左右波动，即 $\hat{\theta}$ 的平均值应该与 θ 的真值相同，

这就是无偏性(没有偏差)的要求。无偏估计的真实意义是：如果相互独立的重复多次用无偏估计量 $\hat{\theta}$ 对 θ 进行估计，那么所得估计值的算术平均值应该与 θ 的真值基本上相同。

当一个估计量不是无偏估计量时，称它为有偏估计量。

样本均值是总体均值的无偏估计量，而样本方差不是总体方差的无偏估计量。

因为 X_i 与总体 X 是独立且同分布的随机变量，故 $E(X_i)=\mu$，$D(X_i)=\sigma^2(i=1, 2, \cdots, n)$，从而

$$E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu$$

所以 $\overline{X}$ 是 μ 的无偏估计量。

而

$$E(S^2)=E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)=\frac{1}{n}\sum_{i=1}^{n}[E(X_i^2)-E(\overline{X}^2)]$$

$$E(X_i^2)=D(X_i)+E(X_i)^2=\sigma^2+\mu^2$$

$$E(\overline{X}^2)=D(\overline{X})+E(\overline{X})^2=\frac{1}{n}\sigma^2+\mu^2$$

所以

$$E(S^2)=\frac{n-1}{n}\sigma^2$$

因此样本方差 S^2 不是 σ^2 的无偏估计量，但是当 $n\to\infty$ 时，$E(S^2)=\sigma^2$，这样的统计量称之为渐近统计量。

实际上，只要对 S^2 作一点修正就可以得到方差的无偏统计量，令修正的样本方差为：

$$S^{*2}=\frac{n}{n-1}S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

则此时

$$E(S^{*2})=E\left(\frac{n}{n-1}S^2\right)=\frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2=\sigma^2$$

所以，修正的样本方差 S^{*2} 即是总体方差的无偏估计量。

无偏估计是点估计的基本要求，它保证 $\hat{\theta}$ 对 θ 的估计只有随机误差，而没有系统误差。

1.3.2.2 一致估计

设 $\hat{\theta}=\hat{\theta}(X_1, X_2, \cdots, X_n)$ 为参数 θ 的一个估计量，n 为样本容量。若对于任意 $\theta\in\Theta$，当 $n\to\infty$ 时，$\hat{\theta}(X_1, X_2, \cdots, X_n)$ 依概率收敛于 θ，则称 $\hat{\theta}$ 为 θ 的一

致估计量。即对任何一个$\varepsilon>0$，式

$$\lim_{n\to\infty}P(|\hat{\theta}_n-\theta|\geqslant\varepsilon)=0 \text{ 或} \lim_{n\to\infty}P(|\hat{\theta}_n-\theta|<\varepsilon)=1$$

成立，则称$\hat{\theta}$为参数θ的一致估计量。

1.3.2.3 有效性

由于θ的无偏估计是不唯一的，那么在θ的无偏估计中哪个更好呢？这里“好”的意思是$\hat{\theta}$的取值更靠近θ或更集中在θ的附近，在统计中常用方差来描述。

设$\hat{\theta}_1$，$\hat{\theta}_2$是总体未知参数θ的两个无偏估计量，若对任意的样本容量n，有方差

$$D(\hat{\theta}_1)\leqslant D(\hat{\theta}_2)$$

且至少对某一个$\theta\in\Theta$，上式中的不等号成立，则称$\hat{\theta}_1$是比$\hat{\theta}_2$有效的估计量。

如果在θ的一切无偏估计量中，$\hat{\theta}_1$的方差最小，则称$\hat{\theta}_1$是θ的有效估计量。

事实上，对有效估计量提出了估计量应该满足波动性小的条件。无系统偏差的估计可能有很多，应该挑选其中波动最小的作为估计量。估计量$\hat{\theta}$的方差$D(\hat{\theta})$是对$\hat{\theta}$波动性的度量，波动最小即要求方差达到最小，这就是有效性的基本要求。

有效性的直观含义是：如果$\hat{\theta}_1$，$\hat{\theta}_2$分布的均值都是θ，若$\hat{\theta}_1$比$\hat{\theta}_2$有效，那么$\hat{\theta}_1$的分布形状比较尖，而$\hat{\theta}_2$的分布形状比较平坦，也就是说，$\hat{\theta}_1$在θ附近取值的概率比$\hat{\theta}_2$大。

1.4 参数假设检验

前面介绍了统计推断的一类方法——未知参数的统计估值，总体未知参数的点估计与区间估计问题。而在一些实际问题中，需要知道总体的未知参数有无明显的变化，或是否达到既定的要求，或多个总体的某个参数有无明显的差异等。下面介绍的统计检验就是统计推断的另一类方法。

我们先看一个例子。如果一个人说他从来没有骗过人，能够证明吗？要证明他没有骗过人，必须出示他从小到大每一时刻的经历，还要证明这些经历是完全的、真实的、没有间断的，这简直是不可能的。即使他找到一些证人，那也只能够证明在那些证人在场的某些片刻，他没有被听到骗过人。反过来，要证明这个人骗过人很容易，只要有一次被抓住就足够了。肯定事物很难，而否定却相对容易得多，这就是假设检验背后的哲学。科学总是在否定中发展。

在假设检验中，一般要设立一个原假设(上面的“从来没骗过人”就是一个例子)，而设立该假设的动机主要是企图利用人们掌握的反映现实世界的数据来

找出假设与现实之间的矛盾，从而否定这个假设。

1.4.1　假设检验的原理

在统计学上，首先是对问题发表“看法”此时称之为假设(hypothesis)，而依据样本用一定的方法论证这一假设是否成立称之为统计检验(statistic test)。对总体的分布函数形式或分布中某些未知参数作出某种假设，然后抽取样本，构造适当的统计量，对假设的正确性进行判断的过程，称为假设检验(何晓群，2003)。

在假设检验中，首先要提出一个原假设，比如某正态总体的均值等于9($\mu=9$)，这种原假设也称为零假设，记为H_0。与此同时必须提出与之对立的假设称为备择假设(或备选假设)，比如总体均值大于9($\mu>9$)，备择假设记为H_1。形式上，这个关于总体均值的H_0相对于H_1的检验记为H_0：$\mu=9$；H_1：$\mu>9$。

在多数统计教材中假设检验都是以否定原假设为目标。如否定不了，说明证据不足，无法否定原假设，但不能说明原假设正确，就像一两次没有听过一个人骗人还远不能证明他从来没有骗过人。

备择假设应该按照现实世界所代表的方向来确定，即它通常是被认为可能比原假设更符合数据所代表的现实，比如上面的H_1为$\mu>9$；这意味着，至少样本均值应该大于9；至于是否显著，依检验结果而定。检验结果显著意味着有理由拒绝原假设。因此，假设检验也被称为显著性检验。

有了两个假设，就要根据数据来对它们进行判断。数据的代表是作为其函数的统计量，它在检验中被称为检验统计量。根据原假设(不是备择假设)，可得到该检验统计量的分布；再看这个统计量的数据实现值属不属于小概率事件。也就是说把数据代入检验统计量，看其值是否落入原假设下的小概率范畴；如果的确是小概率事件，也就是说，原假设发生的概率相对较小，那么就有可能拒绝原假设，或者说“该检验显著”；否则说“没有足够证据拒绝原假设”或者“该检验不显著”。

但小概率并不能说明不会发生，仅仅是发生的概率很小罢了。拒绝正确原假设的错误常被称为第一类错误，其发生的概率称为犯第一类错误的概率，通常记为α，即：

$$P\{\text{拒绝 } H_0 \mid H_0 \text{ 为真}\}\leqslant\alpha$$

在一般情况下，对于给定的α，我们称其为显著性水平，$1-\alpha$称为置信水平。

在备择假设正确时反而说原假设正确所犯错误，称为第二类错误，其发生的概率称为第二类错误的概率，通常记为β，即：

$$P\{\text{接受 } H_0 \mid H_0 \text{ 不真}\}$$

但是在实际的统计推断中，我们大多用犯第一类错误的概率来检验到底是接

受还是拒绝原假设。

对于给定的检验法则，本质上是将样本空间划分为互不相交的两个子集 C，C^*，使得当样本观测点在 C 里面时，拒绝原假设，C 即是拒绝域；而当样本观测点在 C^* 里面时，接受原假设，C^* 即为接受域。

1.4.2 假设检验的步骤

归纳起来，假设检验的步骤为：

1. 提出原假设 H_0 和备择假设 H_1；
2. 给定显著性水平 α 以及样本容量 n；
3. 确定检验统计量以及拒绝域的形式；
4. 按 $P\{拒绝\ H_0 \mid H_0\ 为真\} \leqslant \alpha$ 查出检验统计量临界值，求出拒绝域；
5. 取样，根据样本观察值作出决策，是接受 H_0 还是拒绝 H_0。

大多数假设检验的检验统计量服从 t 分布、χ^2 分布、F 分布或其他特殊的理论分布。进行这类检验时，通常在检验之前确定显著性水平，并从有关表格中查出用于判断的检验统计量临界值，然后将检验统计量与临界值相比，即可作出统计推断。

例如在回归方程的显著性检验中，就是要看自变量 x_1，x_2，…，x_k 从整体上对随机变量 y 是否有明显的影响。为此，可提出假设：

$$H_0：b_1=b_2=\cdots=b_k=0$$

如果 H_0 被接受，则表明随机变量 y 与 x_1，x_2，…，x_k 之间的关系由线性回归模型表示不合适。

有统计量：

$$F=\frac{S_{回}/k}{S_{残}/(n-k-1)}\sim F(k,n-k-1)$$

于是，可利用 F 统计量对回归方程的总体显著性进行检验。对于给定的数据(y，x_{i1}，x_{i2}，…，x_{ik})($i=1$，2，…，n；$n>k+1$)，计算出 $S_{回}$ 和 $S_{残}$，其中 $S_{回}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 为回归平方和，$S_{残}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 为残差平方和，进而得到 F 的值，再由给定的显著性水平 α，查 F 分布表，得临界值 $F_\alpha(k,\ n-k-1)$。

当 $F>F_\alpha(k,\ n-k-1)$，则拒绝假设 H_0，认为在显著性水平 α 下，y 对 x_1，x_2，…，x_k 有显著的线性关系，即回归方程是显著的；若有 $F\leqslant F_\alpha(k,\ n-k-1)$，则没有足够的理由否定 H_0，认为回归方程不显著。其中 $F_\alpha(k,\ n-k-1)$ 是给定显著性水平 α 下，查第一自由度为 k，第二自由度为 $n-k-1$ 的 F 分布表所得的 F 临界值。

1.4.3 参数检验

假设检验在统计模型的显著性检验中具有十分重要的意义。下面对常用的 Z 检验、t 检验、χ^2 检验和 F 检验等参数检验方法作介绍。

1.4.3.1 总体方差 σ^2 已知，检验总体均值 μ

设总体 $X \sim N(\mu, \sigma^2)$，方差为 σ^2，从总体 X 中抽取样本 $X_1, X_2, \cdots, X_n$，样本均值为 $\overline{X}$，检验总体均值 μ，给出如下的三种检验假设：

(1) H_0：$\mu=\mu_0$；H_1：$\mu\neq\mu_0$

(2) H_0：$\mu\geqslant\mu_0$；H_1：$\mu<\mu_0$

(3) H_0：$\mu\leqslant\mu_0$；H_1：$\mu>\mu_0$

其中，μ_0 已知，为方便起见，把(1)中的假设叫做双侧假设；(2)、(3)中的假设叫做单侧假设，对它们所作的检验分别叫做双侧检验和单侧检验。

对于一个正态总体的均值假设检验，当 σ^2 已知时，不论是双侧检验还是单侧检验，都用 $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0, 1)$进行检验，这种用正态变量作为检验统计量的假设检验方法，称为 Z 检验法。表 1.9 列出了单个总体均值的 Z 检验法(陈玉成等，1998)。

表 1.9 σ^2 已知时单个总体均值的 Z 检验法

检验方法	双侧检验	单侧检验	
原假设 H_0	$\mu=\mu_0$	$\mu\geqslant\mu_0$	$\mu\leqslant\mu_0$
备择假设 H_1	$\mu\neq\mu_0$	$\mu<\mu_0$	$\mu>\mu_0$
检验统计量	$Z=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$		
临界值 C	$z_{\alpha/2}$	$-z_\alpha$	z_α
拒绝域	$\lvert Z\rvert\geqslant C$	$Z\leqslant C$	$Z\geqslant C$

例 1.12 检验一污水处理厂出水的氯化物浓度，共采了 25 个水样。设水样的氯化物浓度服从正态分布。问：若标准差为 $\sigma_0=10$ mg/L，水样的氯化物平均浓度为 279 mg/L，出水的氯化物浓度是否超过设计的 250 mg/L 的标准？

解 本问题是检验污水处理厂出水的氯化物浓度是否超过预期设计的标准，这是一个单侧的正态检验。

已知条件：$\mu_0=250$ mg/L，显著性水平 $\alpha=0.05$，$\sigma_0=10$，$n=25$

根据数据，计算得到：$\overline{X}=279$ mg/L

假设检验过程为：

$$H_0: \mu=\mu_0=250$$
$$H_1: \mu>\mu_0$$

检验统计量为：

$$Z=\frac{|\overline{X}-\mu_0|}{\sigma_0/\sqrt{n}}=\frac{|279-250|}{10/\sqrt{25}}=14.5$$

取 $\alpha=0.05$，从附表中查出临界值 $C=z_\alpha=1.645$，显然 $Z>C$，故拒绝原假设 H_0，说明出水的氯化物浓度超过了设计的 250 mg/L 的标准。

1.4.3.2 总体方差 σ^2 未知，检验总体均值 μ

设总体 $X\sim N(\mu, \sigma^2)$，方差 σ^2 未知，从总体 X 中随机抽取样本 X_1，X_2，…，X_n，样本均值与样本方差分别为 $\overline{X}$ 与 S^2，检验总体均值 μ，给出如下三种检验假设：

(1)H_0：$\mu=\mu_0$；H_1：$\mu\neq\mu_0$

(2)H_0：$\mu\geqslant\mu_0$；H_1：$\mu<\mu_0$

(3)H_0：$\mu\leqslant\mu_0$；H_1：$\mu>\mu_0$

对于一个正态总体，方差 σ^2 未知，检验总体均值 μ 类似于方差 σ^2 已知情形的讨论。但是由于总体方差 σ^2 未知，所以统计量 Z 已经不能使用。因为 Z 中含有未知参数 σ^2，它已经不是一个统计量，所以要选取一个不含未知参数 σ^2 的统计量。考虑用方差的渐近无偏估计 S^{*2} 来取代总体方差，这样就得到 t 统计量。

对于一个正态总体的均值假设检验，在总体方差 σ^2 未知的情况下，不论是双侧检验还是单侧检验，都用 $t=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t\ (n-1)$ 进行检验，由于引入的检验统计量均为 t 统计量，故对正态总体均值的检验称为 t 检验法。表 1.10 列出了单个总体均值的 t 检验法。

表 1.10 σ^2 未知时单个总体均值的 t 检验法

检验方法	双侧检验	单侧检验	
原假设 H_0	$\mu=\mu_0$	$\mu\geq\mu_0$	$\mu\leq\mu_0$
备择假设 H_1	$\mu\neq\mu_0$	$\mu<\mu_0$	$\mu>\mu_0$
检验统计量	$t=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$		
临界值 C	$t_{\alpha/2}\ (n-1)$	$-t_{\alpha}\ (n-1)$	$t_{\alpha}\ (n-1)$
统计推断 拒绝域	$\lvert t\rvert>C$	$t<C$	$t>C$

例 1.13 在对 AAS(原子吸收分光光度法)测定淡水沉积物中 Ni 含量的方法进行考核时，使用了已知的 Ni 浓度为 4.55 mg/kg 的参照样。按规定的消解和分析程序对此参照样进行 5 次重复测定，结果(单位：mg/kg)为 4.28，4.40，4.42，4.35，4.37，希望据此判断所使用的测定方法有没有明显的系统误差?

解 该问题用统计语言表述为：总体均值(即用 AAS 进行多次测定所得到数据的均值)与已知值(参照样 4.55 mg/kg)之间有没有显著性差异，属于单个总体均值比较。由于系统误差可能偏高，也可能偏低，故采用双侧 t 检验。

有关已知条件为：$\mu_0=4.55$，$n=5$，$\overline{X}=4.364$，$S=0.054$

则检测过程为：

$$H_0:\ \mu=\mu_0=4.55;\qquad H_1:\ \mu\neq\mu_0$$

检验统计量为：

$$t=\frac{|\overline{X}-\mu_0|}{S/\sqrt{n}}=\frac{|4.364-4.55|}{0.054/\sqrt{5}}=7.702$$

取 $\alpha=0.05$，从附表中查出双侧检验临界值 $C=t_{\alpha/2}(n-1)=t_{\alpha/2}(4)=2.776$，显然 $t>C$，故拒绝 H_0，即这种方法存在明显的系统误差。

1.4.3.3 一个正态总体方差的假设检验

设有正态总体 $X\sim N(\mu,\ \sigma^2)$，μ 与 σ^2 均未知，从总体 X 中随机抽取样本 X_1，X_2，…，X_n，样本方差为 S^2，检验总体方差 σ^2，给出如下三种检验假设：

(1) H_0：$\sigma^2=\sigma_0^2$；H_1：$\sigma^2\neq\sigma_0^2$

(2) H_0：$\sigma^2\geq\sigma_0^2$；H_1：$\sigma^2<\sigma_0^2$

(3) H_0：$\sigma^2\leq\sigma_0^2$；H_1：$\sigma^2>\sigma_0^2$

对于以上三个假设检验问题，在 H_0 为真的条件下，用于检验假设 H_0 的检

验统计量及其分布均为 $\chi^2=\frac{(n-1)\ S^2}{\sigma_0^2}\sim\chi^2(n-1)$。由于引入的检验统计量为 χ^2 统计量，故称对正态总体方差的检验为 χ^2 检验法。表 1.11 列出了单个总体均值的 χ^2 检验法。

表 1.11　单个总体方差的 χ^2 检验法(α 水平)

检验方法	双侧检验	单侧检验	
原假设 H_0	$\sigma^2=\sigma_0^2$	$\sigma^2\geqslant\sigma_0^2$	$\sigma^2\leqslant\sigma_0^2$
备择假设 H_1	$\sigma^2\neq\sigma_0^2$	$\sigma^2<\sigma_0^2$	$\sigma^2>\sigma_0^2$
检验统计量	$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}$		
检验临界值	$\chi^2_{1-\alpha/2}(n-1)$ $\chi^2_{\alpha/2}(n-1)$	$\chi^2_{1-\alpha}(n-1)$	$\chi^2_{\alpha}(n-1)$
拒绝域	$\chi^2>\chi^2_{\alpha/2}\ (n-1)$ 或 $\chi^2<\chi^2_{1-\alpha/2}\ (n-1)$	$\chi^2<\chi^2_{1-\alpha}\ (n-1)$	$\chi^2>\chi^2_{\alpha}\ (n-1)$

例 1.14　一自动车床加工零件的精度服从正态分布 $N(\mu,\ \sigma^2)$，原来加工精度 $\sigma_0^2=0.18$。经过一段时间的生产后，要检验一下这一车床是否保持原来的精度，即检验假设 H_0：$\sigma^2=0.18$。为此抽取这车床所加工的 31 个零件，测得数据如下表（表 1.12）。

表 1.12　测量数据

零件长度 x_i	10.1	10.3	10.6	11.2	11.5	11.8	12.0
频数 n_i	1	3	7	10	6	3	1

在给定显著性水平 $\alpha=0.05$ 的情况下，根据题意只考虑单侧的情形，由

$$P(\chi^2>\chi^2_{0.05}(30))=0.05$$

定出临界值。查自由度为 30 的 χ^2 分布表得 $\chi^2_{0.05}(30)=43.8$。再由样本观察值算出

$$\chi^2=\frac{\sum_{i=1}^{7}n_i(x_i-\bar{x})^2}{0.18}=44.5>43.8=\chi^2_{0.05}(30)$$

因此拒绝原假设 H_0。这说明在显著性水平 $\alpha=0.05$ 下，自动车床工作一段时间后精度变差。

1.4.3.4 两个正态总体参数的假设检验

设两个正态总体 $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$，分别从两个总体中抽取容量为 n_1，n_2 的两个独立样本，计算得样本均值分别为 $\bar{x}_1$ 与 $\bar{x}_2$（为了简便起见，以后的样本均值也用小写字母 $\bar{x}$ 表示），样本方差分别为 S_1^2 与 S_2^2。下面就均值与方差的差异性分别予以讨论。

1. 两个正态总体均值差异性检验

检验目标是两个总体均值的差异性，与一个总体假设类似，作出双侧假设和单侧假设如下：

(1) H_0：$\mu_1 = \mu_2$；H_1：$\mu_1 \neq \mu_2$

(2) H_0：$\mu_1 \geqslant \mu_2$；H_1：$\mu_1 < \mu_2$

(3) H_0：$\mu_1 \leqslant \mu_2$；H_1：$\mu_1 > \mu_2$

对于两个正态总体均值的假设检验，表 1.13 列出了当 σ_1^2，σ_2^2 均已知时及 σ_1^2，σ_2^2 均未知时的检验法。

表 1.13　两个独立总体均值比较的 Z 检验和 t 检验（α 水平）

检验方法	Z 检验		t 检验	
适用情景	σ_1^2，σ_2^2 已知		σ_1^2，σ_2^2 未知	
原假设 H_0	$\mu_1 = \mu_2$	$\mu_1 \leqslant \mu_2$ 或 $\mu_1 \geqslant \mu_2$	$\mu_1 = \mu_2$	$\mu_1 \leqslant \mu_2$ 或 $\mu_1 \geqslant \mu_2$
备择假设 H_1	$\mu_1 \neq \mu_2$	$\mu_1 > \mu_2$ 或 $\mu_1 < \mu_2$	$\mu_1 \neq \mu_2$	$\mu_1 > \mu_2$ 或 $\mu_1 < \mu_2$
检验统计量	$Z = \dfrac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$		$t = \dfrac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	
检验临界值 C	$z_{\alpha/2}$	z_{α}	$t_{\alpha/2}(n_1+n_2-2)$	$t_{\alpha}(n_1+n_2-2)$

例 1.15　已知放射强度服从正态分布。对甲、乙两个放射污染区进行反射强度测定，从甲地取得样本数为 63，其结果符合 $N_1(62.3, 10.8)$；从乙地取得样本数为 74，其结果符合正态分布 $N_2(66.8, 13.3)$。问甲、乙两地放射污染强度是否相同?

解　依题意，该问题属于两个独立总体均值比较的假设检验，且变量服从正态分布，σ_1^2，σ_2^2 均已知，故采用正态 Z 检验中的双侧检验。

有关已知条件为：$n_1 = 63$，$\bar{x}_1 = 62.3$，$\sigma_1^2 = 10.8$；$n_2 = 74$，$\bar{x}_2 = 66.8$，$\sigma_2^2 = 13.3$

则检测过程为：

$$H_0: \mu_1=\mu_2; \quad H_1: \mu_1 \neq \mu_2$$

检验统计量为：

$$Z=\frac{|\bar{x}_1-\bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{|62.3-66.8|}{\sqrt{\frac{10.8}{63}+\frac{13.3}{74}}}=7.59$$

取 $\alpha=0.05$，从附表中查出双侧检验临界值 $C=z_{\alpha/2}=1.960$，显然 $Z>C$，故拒绝 H_0，则甲、乙两地放射污染强度显著不同。

2. *两个正态总体方差的差异性检验*

假设两个正态总体均值 μ_1，μ_2 未知，检验两个总体方差的差异性。其假设如下：

(1) H_0：$\sigma_1^2=\sigma_2^2$；H_1：$\sigma_1^2 \neq \sigma_2^2$

(2) H_0：$\sigma_1^2 \leqslant \sigma_2^2$；$H_1$：$\sigma_1^2 > \sigma_2^2$

(3) H_0：$\sigma_1^2 \geqslant \sigma_2^2$；$H_1$：$\sigma_1^2 < \sigma_2^2$

对于上述三种统计假设的检验，可以采用表 1.14 给出的 F 检验法。

表 1.14　两个总体方差的 F 检验(α 水平)

检验方法		双侧检验	单侧检验	
原假设 H_0		$\sigma_1^2=\sigma_2^2$	$\sigma_1^2 \leqslant \sigma_2^2$	$\sigma_1^2 \geqslant \sigma_2^2$
备择假设 H_1		$\sigma_1^2 \neq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$
检验统计量	F	$F=S_1^2/S_2^2$	$F=S_1^2/S_2^2$	$F=S_1^2/S_2^2$
	v_1	n_1-1	n_1-1	n_1-1
	v_2	n_2-1	n_2-1	n_2-1
检验临界值		$F_{\alpha/2}(n_1-1, n_2-1)$， $F_{1-\alpha/2}(n_1-1, n_2-1)$	$F_\alpha(n_1-1, n_2-1)$	$F_{1-\alpha}(n_1-1, n_2-1)$
拒绝域		$F>F_{\alpha/2}$ 或 $F<F_{1-\alpha/2}$	$F>F_\alpha$	$F<F_{1-\alpha}$

例 1.16　两种型号脱硫装置的脱硫效率十分接近，进行 3 次重复试验后，得表 1.15，根据表中数据，对两种型号装置的稳定性进行比较。

表 1.15　脱硫装置的脱硫效率

型号 1	98	82	96
型号 2	92	95	89

解 依题意为两个总体的方差检验，有关条件已知为：$n_1=3$，$S_1^2=76$；$n_2=3$，$S_2^2=9$。

检验过程为：

$$H_0: \sigma_1^2=\sigma_2^2; \ H_1: \sigma_1^2\neq\sigma_2^2$$

由于：$S_1^2>S_2^2$，则检验统计量

$$F=\frac{S_1^2}{S_2^2}=76/9=8.44$$

双侧检验临界值：$F_{0.025}(2, 2)=39.00$，因此不能拒绝 H_0，结论为在 0.05 显著性水平下未发现两种设备的运转稳定性有明显差别。

1.4.3.5 总结

在确定了参数检验之后，针对所研究的问题本身应选择具体的假设检验方法。常用的假设检验方法总结如下表。

表 1.16 正态总体参数的显著性假设检验

检验参数		假设 H_0	统计量	分布
单个总体	μ	$\mu=\mu_0$(σ^2 已知)	$Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$N(0, 1)$
		$\mu=\mu_0$(σ^2 未知)	$t=\dfrac{\bar{x}-\mu_0}{S/\sqrt{n}}$	$t(n-1)$
	σ^2	$\sigma^2=\sigma_0^2$(μ 已知)	$\chi^2=\dfrac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sigma_0^2}$	$\chi^2(n)$
		$\sigma^2=\sigma_0^2$(μ 未知)	$\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$
两个总体	μ	$\mu_1=\mu_2$(σ_1^2，σ_2^2 已知)	$Z=\dfrac{\lvert\bar{x}_1-\bar{x}_2\rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$N(0, 1)$
		$\mu_1=\mu_2$(σ_1^2，σ_2^2 未知)	$t=\sqrt{\dfrac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}\cdot\dfrac{\lvert\bar{x}_1-\bar{x}_2\rvert}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}}$	$t(n_1+n_2-2)$
	σ^2	$\sigma_1^2=\sigma_2^2$(μ_1，μ_2 未知)	$F=\dfrac{S_1^2}{S_2^2}$	$F(n_1-1, n_2-1)$

1.5 方差分析与试验设计初步

方差分析(analysis of variance，ANOVA)由英国大统计学家费歇尔在20世纪20年代创立的。当时他在英国一个农业站工作，需要进行许多田间试验，为分析试验结果，他发明了方差分析法(于义良，2002)。为纪念费歇尔，方差分析又称 F 检验。后来ANOVA被广泛应用于分析心理学、生物学、环境科学与环境工程和医药等试验数据的分析。从形式上看，方差分析是比较多个总体均值是否相等，但本质上是研究变量之间的关系。方差分析与回归分析有许多相同之处，但又存在本质区别，方差分析主要研究分类型变量对数值型变量的影响，比如它们之间有没有关系、关系的强度如何等；而回归分析主要研究数值型自变量和数值型因变量之间的关系。

1.5.1 方差分析概述

首先从一个例子说起：

例1.17 某公司研究三种内容的广告宣传对某种环境产品销售量的影响，他们对其进行了调查统计。经广告以不同的内容广泛宣传后，按寄回的广告上的订购数计算，一年四个季度的销售量情况如下表(杨虎等，2006)。

表1.17　某环境产品销售量数据表

广告类型	季度				$\bar{x}$
	一	二	三	四	
A_1	163	176	170	185	173
A_2	184	198	179	190	188
A_3	206	191	218	224	210

表中，广告 A_1 强调运输的方便性、A_2 强调节省燃料的经济性、A_3 强调噪声低的优良性，试判断广告的类型对该种环境产品的销售量是否有显著的影响？若有影响，哪种广告内容比较好？

判断广告的类型对环境产品销售量是否有显著的影响，作出这种判断最终被归结为检验这三种广告内容下的环境产品销售量的均值是否相等。如果它们相等，就意味着“广告类型”对销售量没有影响，也就是各种广告下销售量没有显

著差异；如果均值不相等，则意味着广告类型对销售量是有影响的。

为了方便表述，我们作如下定义：在方差分析中，所要检验的对象称为因素；因素的不同表现称为水平或处理；每个因素水平下得到的样本数据称为观测值。

上例中，广告为因素，广告的三种类型称为水平，每个广告类型下的样本数据为观测值。我们怎样判断广告对销售量有显著影响呢？容易想到，如果广告类型对销售量没有影响，那么三个样本可以认为来自同一个总体 $N(\mu, \sigma^2)$，此处 $\mu=\mu_1=\mu_2=\mu_3$，反之，如果广告类型对销售量有影响，则 μ_1，μ_2，μ_3 有显著差异，因此，我们把问题转化为用三个样本去检验假设：

$$H_0: \mu_1=\mu_2=\mu_3; \quad H_1: \mu_1, \mu_2, \mu_3 \text{ 不全相等}$$

从表中看出，各样本均值之间确实存在差异，那么是否可以说明广告对销售量有显著影响呢？不能，因为同一水平(同广告类型)下各试验数据之间还有差异，这显然是由广告类型以外的其他随机影响引起的随机误差，它也会引起各样本均值之间的差异。因此，问题不在于各种样本均值之间是否有差异，而在于这种差异与随机误差相比是否显著偏大，如果是，就有理由认为广告类型对销售量有显著影响，从而否定 H_0。

1.5.2　单因素方差分析

当方差分析只涉及一个分类型自变量时，称为单因素方差分析。例如检验不同广告销售量是否相等，这里只涉及“广告类型”一个因素，也就是单因素方差分析。

1.5.2.1　数据结构

进行单因素方差分析时，需要有下面的数据结构，如表 1.18 所示。

表 1.18　　单因素方差分析的数据结构

水平	观测值			
	1	2	…	n_i
A_1	x_{11}	x_{12}	…	x_{1n_1}
A_2	x_{21}	x_{22}	…	x_{2n_2}
⋮	⋮	⋮	⋮	⋮
A_k	x_{k1}	x_{k2}	…	x_{kn_k}

在单因素分析中，用 A 表示因素，因素的 k 个水平（总体）分别用 A_1，A_2，…，A_k 表示，每个观测值用 x_{ij} 表示，从不同水平中所抽取的样本容量，可以相等也可以不等。

1.5.2.2 分析步骤

方差分析步骤（贾俊平，2005）：

1. 提出假设

检验因素的 k 个水平的均值是否相等，提出如下形式的假设：

H_0：$\mu_1=\mu_2=\cdots=\mu_i=\cdots=\mu_k$；$H_1$：$\mu_1$，$\mu_2$，…，$\mu_i$，…，$\mu_k$ 不全相等

需要注意的是，拒绝原假设 H_0 时，只是表明至少有两个总体的均值不相等，并不意味着所有的均值都不相等。

2. 构造检验的统计量

（1）计算因素各水平的均值。

假定从第 i 个因素水平总体中抽取一个容量为 n_i 的样本，令 $\bar{x}_i$ 为第 i 个因素总体的样本均值，则有：

$$\bar{x}_i=\frac{\sum_{j=1}^{n_i}x_{ij}}{n_i} \qquad (i=1，2，\cdots，k)$$

式中，n_i 为第 i 个因素水平总体的样本观测个数，x_{ij} 为第 i 个因素水平总体的第 j 个观测值。

（2）计算全部观测值的总平均值。

$$\bar{\bar{x}}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}}{n}=\frac{\sum_{i=1}^{k}n_i\bar{x}_i}{n} \qquad \left(n=\sum_{i=1}^{k}n_i\right)$$

（3）计算误差平方和。为构造检验统计量，在方差分析中，需要计算 3 个误差平方和，它们分别是总偏差平方和、水平项误差平方和以及误差平方和。

①总偏差平方和，简记为 SS_T，它是全部观测值 x_{ij} 与总平均值 $\bar{\bar{x}}$ 的误差平方和，反映了全部试验数据之间的差异，因此 SS_T 又称为总变差。其计算公式为：

$$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2 \tag{1.28}$$

②水平项误差平方和，简记为 SS_A，它是各组平均值 $\bar{x}_i(i=1，2，\cdots，k)$ 与总平均值 $\bar{\bar{x}}$ 的误差平方和，反映各总体的样本均值之间的差异程度，因此又称为组间平方和，其计算公式为：

$$SS_A = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{\bar{x}})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2 \tag{1.29}$$

③误差平方和，简记为 SS_E，它是每个水平的各样本数据与其组平均值误差的平方和，反映了每个样本各观测值的离散状况，因此又称为组内平方和或残差平方和，该平方和反映的是随机误差的大小，其计算公式为：

$$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \tag{1.30}$$

上述三个平方和之间的关系为：

$$SS_T = SS_A + SS_E \tag{1.31}$$

④计算统计量

我们知道，各误差平方和的大小与观测值的多少有关。为了消除观测值多少对误差平方和大小的影响，需要将其平均，也就是用各平方和除以它们对应的自由度，这一结果称为均方。三个平方和对应的自由度分别为：SS_T 的自由度为 $n-1$，其中 n 为全部观测值的个数；SS_A 的自由度为 $k-1$，其中 k 为因素水平的个数；SS_E 的自由度为 $n-k$。

SS_A 的均方记为 MS_A，其计算公式为：

$$MS_A = \frac{SS_A}{k-1} \tag{1.32}$$

SS_E 的均方记为 MS_E，其计算公式为：

$$MS_E = \frac{SS_E}{n-k} \tag{1.33}$$

将上述的 MS_A 和 MS_E 进行对比，即得到所需要的检验统计量 F。当 H_0 为真时，二者的比值服从分子自由度为 $k-1$，分母自由度为 $n-k$ 的 F 分布，即：

$$F = \frac{MS_A}{MS_E} \sim F(k-1,\ n-k) \tag{1.34}$$

3. 统计决策

计算出检验的统计量后，将统计量的值 F 与给定的显著性水平 α 的临界值 F_α 进行比较，从而作出对原假设 H_0 的决策。

根据给定的显著性水平 α，在 F 分布表中查找与分子自由度 $df_1=k-1$、分母自由度 $df_2=n-k$ 相应的临界值 $F_\alpha(k-1,\ n-k)$。若 $F>F_\alpha$，则拒绝原假设 H_0，即 $\mu_1=\mu_2=\cdots=\mu_i=\cdots=\mu_k$ 不成立，表明 $\mu_i(i=1,\ 2,\ \cdots,\ k)$之间的差异是显著的；若 $F<F_\alpha$，则不拒绝原假设 H_0，不能认为 $\mu_i(i=1,\ 2,\ \cdots,\ k)$之间有显著差异。

表 1.19　　单因素方差分析表

方差来源	平方和	自由度	均方	F 值	F 临界值
组间	SS_A	$k-1$	$MS_A=\frac{SS_A}{k-1}$	$F=\frac{MS_A}{MS_E}$	$F_\alpha(k-1,\ n-k)$
组内	SS_E	$n-k$	$MS_E=\frac{SS_E}{n-k}$		
总和	SS_T	$n-1$			

1.5.2.3　案例

例 1.18　为比较几种不同类型隔音材料对噪声的去除效果，分别在相同条件下进行若干次重复试验，结果如表 1.20，问不同材料对噪声衰减率是否有明显影响？

表 1.20　　不同类型隔音材料的噪声去除效果

材料	噪声衰减率				样本量
A_1	0.140	0.142	0.144		3
A_2	0.152	0.150	0.156	0.154	4
A_3	0.160	0.158	0.163	0.161	4
A_4	0.175	0.173			2
A_5	0.180	0.184	0.182	0.186	4

解　这是一个单因素试验，其水平数为 5，总样本数为 17，其统计假设为：H_0：5 种材料的隔音效果无明显差异；H_1：5 种材料的隔音效果有明显差异将有关统计量列入表 1.21 中。

表 1.21　　不同类型隔音材料的样本统计量

材料	A_1	A_2	A_3	A_4	A_5	合计
n_i	3	4	4	2	4	17
$\bar{x}_i$	0.142	0.153	0.161	0.174	0.183	0.162

$SS_T=0.003\ 646$

$SS_A=0.003\ 583$

$SS_E = SS_T - SS_A = 0.000\ 063$

方差分析过程见表1.22。

表1.22　　不同类型隔音材料噪声去除效果的方差分析

方差来源	平方和	自由度	均方	F值	$F_{0.05}$	$F_{0.01}$
组间	0.003 583	4	0.000 896	170.61	3.26	5.41
组内	0.000 063	12	0.000 005 25			
总计	0.003 646	16				

由表1.22可以看出，对于给定的显著性水平$\alpha=0.05$或0.01，不同材料对噪声衰减率都有明显影响。

1.5.3 双因素方差分析

在许多实际问题中，往往需要考虑几个因素对试验结果的影响，例如，对环境产品销售量的影响因素不仅有广告，可能还有销售价格等因素。双(多)因素方差分析方法就是研究两种(多种)因素对试验指标的影响程度的分析方法。

由于存在两个因素对试验指标的影响，各个因素不同水平的搭配可能对试验指标产生新的影响，这种现象在统计上称为交互效应。如关于“男性的肥胖比女性的肥胖更容易引起高血压”这种说法，描述的是超重状态下的血压与性别有关，反映了体重、性别对血压可能产生交互效应。各因素是否存在交互效应是多因素方差分析产生的新问题，反映了单因素方差分析与多因素方差分析的本质区别，本书分两种情况进行讨论：一种是无交互作用的双因素方差分析，另一种是有交互作用的双因素方差分析。

1.5.3.1 无交互作用的双因素方差分析

当方差分析中涉及两种类型自变量时，称为双因素方差分析。

1. *数据结构*

无交互作用的双因素方差分析的数据结构，如表1.23所示。由于有两个因素，因此其中一个因素安排在“行”的位置，称为行因素；另一个因素安排在“列”的位置，称为列因素。我们设行因素有k个水平：行1，行2，…，行k；列因素有r个水平：列1，列2，…，列r。行因素和列因素的每一水平都可以搭配成一组，观察它们对试验指标的影响，共抽取$k \times r$个观测数据。每一个观测

值 x_{ij}($i=1, 2, \cdots, k$; $j=1, 2, \cdots, r$)看作是由行因素的 k 个水平和列因素的 r 个水平所组成的 $k\times r$ 个总体中抽取的容量为 1 的独立随机样本。这 kr 个总体中的每一个总体都服从正态分布，且有相同的方差。

表 1.23　　双因素方差分析的数据结构

行因素(i)	列因素(j)				平均值 $\bar{x}_{i.}$
	列 1	列 1	…	列 r	
行 1	x_{11}	x_{12}	…	x_{1r}	$\bar{x}_{1.}$
行 2	x_{21}	x_{22}	…	x_{2r}	$\bar{x}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
行 k	x_{k1}	x_{k2}	…	x_{kr}	$\bar{x}_{k.}$
平均值 $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	…	$\bar{x}_{.r}$	$\bar{x}$

其中 $\bar{x}_{i.}$ 是行因素的第 i 个水平下各观测值的平均值，其计算公式为：

$$\bar{x}_{i.}=\frac{\sum_{j=1}^{r}x_{ij}}{r}\qquad (i=1, 2, \cdots, k)$$

$\bar{x}_{.j}$ 是列因素的第 j 个水平下各观测值的平均值，其计算公式为：

$$\bar{x}_{.j}=\frac{\sum_{i=1}^{k}x_{ij}}{k}\qquad (j=1, 2, \cdots, k)$$

$\bar{x}$ 是全部 kr 个样本数据的总平均值，其计算公式为：

$$\bar{x}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{r}x_{ij}}{kr}$$

2. 分析步骤

与单因素分析类似，双因素方差分析也包括提出假设、确定检验的统计量、决策分析等步骤。

(1)提出假设

为了检验两个因素的影响，需要对两个因素分别提出如下假设：

对行因素提出的假设为：

H_0：$u_1=u_2=\cdots=u_i=\cdots=u_k$；$H_1$：$u_i$($i=1, 2, \cdots, k$)不全相等

对列因素提出的假设为：

H_0：$\mu_1=\mu_2=\cdots=\mu_j=\cdots=\mu_r$；$H_1$：$\mu_j(j=1, 2, \cdots, r)$不全相等

(2)构造检验统计量

为了检验 H_0 是否成立，我们需要分别确定检验行因素和列因素的统计量。与单因素方差分析构造统计量的方法一样，也需要从总误差平方和的分解入手。总偏差平方和是全部样本观测值 $x_{ij}(i=1, 2, \cdots, k; j=1, 2, \cdots, r)$与总的样本平均值的误差平方和 $\bar{\bar{x}}$，记为 SS_T，即：

$$\begin{aligned} SS_T &= \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{\bar{x}})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{i.}-\bar{\bar{x}})^2+\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{.j}-\bar{\bar{x}})^2+\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}})^2 \\ &= r\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{\bar{x}})^2+k\sum_{j=1}^{r}(\bar{x}_{.j}-\bar{\bar{x}})^2+\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}})^2 \end{aligned} \tag{1.35}$$

其中，分解后的等式右边第一项是行因素所产生的误差平方和，记为 SS_R，即：

$$SS_R = r\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{\bar{x}})^2 \tag{1.36}$$

第二项是列因素所产生的误差平方和，记为 SS_C，即：

$$SS_C = k\sum_{j=1}^{r}(\bar{x}_{.j}-\bar{\bar{x}})^2 \tag{1.37}$$

第三项是除行因素和列因素之外的剩余因素影响产生的误差平方和，称为随机误差平方和，记为 SS_E，即：

$$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}})^2 \tag{1.38}$$

上述各平方和的关系为：

$$SS_T=SS_R+SS_C+SS_E \tag{1.39}$$

在上述误差平方和的基础上，计算均方，其中与各误差平方和相对应的自由度分别是：总误差平方和 SS_T 的自由度为 $kr-1$；行因素的误差平方和 SS_R 的自由度为 $k-1$；列因素的误差平方和 SS_C 的自由度为 $r-1$；随机误差平方和 SS_E 的自由度为$(k-1)(r-1)$。

为了构造检验统计量，需要计算下列各均方：

行因素的均方，记为 MS_R：

$$MS_R=\frac{SS_R}{k-1} \tag{1.40}$$

列因素的均方，记为 MS_C：

$$MS_C=\frac{SS_C}{r-1} \tag{1.41}$$

随机误差的均方，记为 MS_E：

$$MS_E=\frac{SS_E}{(k-1)(r-1)} \tag{1.42}$$

为了检验行因素对因变量的影响是否显著，采用下面的统计量：

$$F_R=\frac{MS_R}{MS_E}\sim F(k-1,\ (k-1)(r-1)) \tag{1.43}$$

为了检验列因素的影响是否显著，采用下面的统计量：

$$F_C=\frac{MS_C}{MS_E}\sim F(r-1,\ (k-1)(r-1)) \tag{1.44}$$

(3)统计决策

计算出检验统计量后，给定的显著性水平 α 和两个自由度，查 F 分布表得到相应的临界值 F_α，然后将 F_R，F_C 与 F_α 进行比较：

若 $F_R>F_\alpha$，则拒绝原假设 H_0，即 $u_1=u_2=\cdots=u_i=\cdots=u_k$ 不成立，表明它们之间的差异是显著的。

若 $F_C>F_\alpha$，则拒绝原假设 H_0，即 $\mu_1=\mu_2=\cdots=\mu_j=\cdots=\mu_r$ 不成立，表明它们之间的差异是显著的。

表 1.24　双因素方差分析表的结构

方差来源	平方和	自由度	均方	F 值	F 的临界值
行因素 A	SS_R	$k-1$	$MS_R=\frac{SS_R}{k-1}$	$F_R=\frac{MS_R}{MS_E}$	$F_\alpha(k-1,\ (k-1)(r-1))$
列因素 B	SS_C	$r-1$	$MS_C=\frac{SS_C}{r-1}$	$F_C=\frac{MS_C}{MS_E}$	$F_\alpha(r-1,\ (k-1)(r-1))$
误差	SS_E	$(k-1)(r-1)$	$MS_E=\frac{SS_E}{(k-1)(r-1)}$		
总和	SS_T	$kr-1$			

3. 案例

例 1.19　为了提高某种环保产品的合格率，考察原料用量和来源地对产品合格率是否有影响，假设原料来源于三个地方：甲、乙、丙，原料的使用量有三种方案：现用量、增加 5%、增加 8%。每个水平组合各作一次试验，得到表 1.25的数据，试分析原料用量及来源地对产品合格率的影响是否显著？

表 1.25 数据表

原料来源地(A)	环保产品合格率		
	B_1(现用量)	B_2(增加 5%)	B_3(增加 8%)
甲(A_1)	59	70	66
乙(A_2)	63	74	70
丙(A_3)	61	66	71

解 设有两个因素 A，B，它们分别对应于产品的来源地和原料用量，显然因素 A 有三个水平 A_1，A_2，A_3，因素 B 也有三个水平 B_1，B_2，B_3，因为各组 (A_i, B_j)中只采样一个数据，组分这种情况下没有交互效应。采用双因素方差分析，得到方差分析，见表 1.26。

表 1.26 双因素方差分析表

方差来源	平方和	自由度	均方	F 值	$F_{0.05}(2, 4)$
因素 A	26	2	13	1.86	6.94
因素 B	146	2	73	10.43	6.94
误差	28	4	7		
总和	200	8			

$F_A=1.86<F_{0.05}(2, 4)=6.94$，$F_B=10.43>F_{0.05}(2, 4)=6.94$，即根据现有数据资料，有 95%的把握推断原料来源地对产品的合格率影响不大，而原料使用量对合格率有显著影响。

1.5.3.2 有交互作用的双因素方差分析

在上面的分析中，我们假定两个因素对因变量的影响是独立的，但如果两个因素搭配会对因变量产生一种新的效应，就需要考虑交互作用对因变量的影响，这就是有交互作用的双因素方差分析。

有交互作用的双因素方差分析也需要提出假设、构造检验的统计量、决策分析等步骤。方法与上述类似，有交互作用的双因素试验数据，见表 1.27。

表 1.27　有交互作用的双因素试验数据表

行因素(i)	列因素(j) 列 1	列 2	…	列 r	平均值 $\bar{x}_{i.}$
行 1	x_{111}，…，$x_{11m_{11}}$	x_{121}，…，$x_{12m_{12}}$	…	x_{1r}，…，$x_{1m_{1r}}$	$\bar{x}_{1.}$
行 2	x_{211}，…，$x_{21m_{21}}$	x_{221}，…，$x_{22m_{22}}$	…	x_{2r}，…，$x_{2m_{2r}}$	$\bar{x}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
行 k	x_{k11}，…，$x_{k1m_{k1}}$	x_{k21}，…，$x_{k2m_{k2}}$	…	x_{kr}，…，$x_{krm_{kr}}$	$\bar{x}_{k.}$
平均值 $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	…	$\bar{x}_{.r}$	$\bar{\bar{x}}$

其中，设行变量有 k 个水平，列变量有 r 个水平，每个水平交叉构成一个样本，每行的样本容量合计为 $m_{i.}$，每列的样本容量合计为 $m_{.j}$；x_{ijl} 为对应于行因素的第 i 个水平和列因素的第 j 个水平的第 l 列的观测值；$\bar{x}_{i.}$ 为行因素的第 i 个水平的样本均值；$\bar{x}_{.j}$ 为列因素的第 j 个水平的样本均值；$\bar{x}_{ij}$ 为对应于行因素的第 i 个水平和列因素的第 j 个水平组合的样本均值；$\bar{\bar{x}}$ 为全部 n 个观测值的总均值，$n=\sum_{i=1}^{k}\sum_{j=1}^{r}m_{ij}$。

各平方和的计算公式如下：

总偏差平方和(SS_T)：

$$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m_{ij}}(x_{ijl}-\bar{\bar{x}})^2 \tag{1.45}$$

行变量平方和(SS_R)：

$$SS_R=\sum_{i=1}^{k}m_{i.}(\bar{x}_{i.}-\bar{\bar{x}})^2 \tag{1.46}$$

列变量平方和(SS_C)：

$$SS_C=\sum_{j=1}^{r}m_{.j}(\bar{x}_{.j}-\bar{\bar{x}})^2 \tag{1.47}$$

交互作用平方和(SS_{RC})：

$$SS_{RC}=\sum_{i=1}^{k}\sum_{j=1}^{r}m_{ij}(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}})^2 \tag{1.48}$$

误差平方和(SS_E)：

$$SS_E=SS_T-SS_R-SS_C-SS_{RC} \tag{1.49}$$

表 1.28 有交互作用的双因素方差分析表的结构

方差来源	平方和	自由度	均方	F值	F的临界值
行因素	SS_R	$k-1$	$MS_R=\frac{SS_R}{k-1}$	$F_R=\frac{MS_R}{MS_E}$	$F_\alpha(k-1, n-kr)$
列因素	SS_C	$r-1$	$MS_C=\frac{SS_C}{r-1}$	$F_C=\frac{MS_C}{MS_E}$	$F_\alpha(r-1, n-kr)$
交互作用	SS_{RC}	$(k-1)(r-1)$	$MS_{RC}=\frac{SS_{RC}}{(k-1)(r-1)}$	$F_{RC}=\frac{MS_{RC}}{MS_E}$	$F_\alpha((k-1)(r-1), n-kr)$
误差	SS_E	$n-kr$	$MS_E=\frac{SS_E}{n-kr}$		
总和	SS_T	$n-1$			

1.5.4 试验设计初步

试验设计已成为数理统计的一个重要分支，其数据分析方法主要是方差分析，本书主要介绍试验设计的一些基本知识。

1.5.4.1 完全随机化设计

收集样本数据的过程称为试验。收集样本数据的计划称为试验设计。试验设计研究如何科学地安排试验，使我们能用尽可能少的试验获得尽可能多的信息。把一批实验对象完全混合，然后分成若干组，再把各因素不同水平的组合作为"处理"随机地安排在这些组上，称为完全随机化设计。接受"处理"的对象或实体，称为试验单元或抽样单元。

例 1.20 一家环境保护公司利用氨基酸产品对废水进行回收利用，研究出一种氨基酸复合肥，公司需要研究不同肥料对小麦产量的影响，为此选择了无机肥、普通有机肥和氨基酸复合肥进行比较，需要选择一些地块，在每个地块施等量的肥料，然后获得产量数据，进而分析肥料对产量的影响是否显著，这一过程就是试验设计的过程。

这里的"肥料种类"就是试验因子或因素，无机肥、普通有机肥和氨基酸复合肥就是因子的三个不同水平，我们称为处理。假定我们选取 3 个面积相同的地块，这里的地块就是接受处理的对象或实体，称为试验单元，然后将每个品种随机地指派给其中的一个地块，例如无机肥可以随机地指派给地块 2，普通有机肥可以指派给地块 1，氨基酸复合肥指派给地块 3，这一过程就是随机化设计过程。

完全随机化设计除符合"随机化"过程外，还必须符合"可重复性"原则，

重复是指在一个试验中每个试验条件可以“复制”，例如在上面例子中，由于只抽取了 3 个地块，只能获得 3 个产量数据，对应于每个处理的样本容量为 1。为获得更多的数据，必须重复基本试验步骤，例如抽取 12 个地块，将每个处理之一随机地指派给其中的 3 个地块，这就相当于重复做了 4 次试验。

假定我们通过上述设计后得到了如下样本数据，见表 1.29。

表 1.29　　3 种肥料在 12 个地块上的产量数据

类　型	产　量			
无机肥	368	349	351	342
普通有机肥	386	383	370	357
氨基酸复合肥	351	348	336	331

要分析肥料类型对产量是否有显著影响，我们用上面介绍的单因素方差分析方法进行分析。表 1.30 给出了对数据的分析结果。

表 1.30　　3 种肥料的方差分析表

方差来源	平方和	自由度	均方	F 值	$F_{0.05}$
组间	2 186	2	1 093	8.42	4.26
组内	1 168	9	130		
总和	3 354	11			

由表 1.30 的计算结果可知，$F>F_{0.05}$，表明肥料种类对产量有显著影响。

1.5.4.2　因子设计

假定除了关心肥料类型对产量的影响外，我们还关心小麦品种对产量的影响。这时我们感兴趣的因素有两个，为肥料类型和小麦品种。假定有甲、乙两种小麦，这样 3 种肥料和 2 种小麦的搭配共有 $3\times2=6$ 种。如果我们选择 30 个地块进行试验，每一种搭配可以做 5 次实验，也就是每个肥料种类的样本容量为 5。这种考虑两个因素(可推广到多个因素)的搭配试验设计称为因子设计。

例 1.21　假定对 3 种肥料，2 个品种小麦的因子试验取得了下面的数据，见表 1.31。

现在我们需要分析小麦品种、肥料类型以及两者交互作用对产量的影响。采用 Excel 中的“可重复双因素分析”得到下面的输出结果，见表 1.32。

由于检验肥料类型的 $F_R > F_{0.05}$，表明肥料类型对产量有显著影响；检验小麦品种的 $F_C > F_{0.05}$，表明小麦品种对产量有显著影响；检验交互作用的 $F_{RC} < F_{0.05}$，表明不能认为肥料类型和小麦品种的交互作用对产量有显著影响。

表 1.31　　肥料类型和小麦品种的因子试验数据

肥料类型	小麦品种	
	甲	乙
无机肥	81	89
	82	92
	79	87
	81	85
	78	86
普通有机肥	71	77
	72	81
	72	77
	66	73
	72	79
氨基酸复合肥	76	89
	79	87
	77	84
	76	87
	78	87

表 1.32　　小麦品种和肥料类型因子试验的方差分析表

方差来源	平方和	自由度	均方	F 值	F_α
行因素	560	2	280	54.37	3.40
列因素	480	1	480	93.20	4.26
交互作用	10.4	2	5.2	1.01	3.40
误差	123.6	24	5.15		
总和	1 174	29			

从前面讨论的单因素和双因素试验均需把每个因素的各种水平相互搭配逐一进行，这对多因素试验来说，将意味着耗费大量的人力、物力、财力和时间。比如，5 个因素，每个因素取 4 个水平，一一搭配需做 $4^5 = 1\ 024$ 次试验，通常这是实际

情况所不允许的。因此，对于多因素的试验，有一个科学安排试验的问题。试验安排得好，既可以减少试验次数，又能获得有效的结果。正交试验设计就是一种合理安排多因素试验的科学方法。有兴趣了解的同学可以查阅相关的文献资料。

【思考题 1】

1. 举例说明常用的几个统计量。

2. 试述区间估计的意义，说明区间估计中显著性水平、置信区间、区间大小、置信系数大小的意义。

3. 详述假设检验的步骤。

4. 试述频率直方图和累积频率图的步骤。

5. 试述正态分布、χ^2 分布、t 分布、F 分布之间的关系，并提供相关数学表达式的相互转换。

6. 已知某地水体 COD 浓度 $X \sim N(2, 3^2)$，求 COD 浓度落在区间(3，9)的概率。

7. 设 X_1，X_2，…，X_n 是取自总体 X 的样本，总体期望 $E(X)=\mu$ 未知，a_1，a_2，…，a_n 为常数，且 $a_1+a_2+\cdots+a_n=1$，求证：$\sum_{i=1}^{n} a_i X_i$ 为 $E(X)=\mu$ 的一个无偏估计。

8. 一台包装机装净水剂，额定标准重量为 500 g。根据以往经验，包装机实际装袋重量服从正态分布 $N(\mu_0, \sigma_0^2)$，其中 $\sigma_0=15$ g，为检验包装机工作是否正常，随机抽取 9 袋，称得净水剂重量(单位：g)数据如下：

497，506，518，524，488，517，510，515，516

若取显著性水平 $\alpha=0.01$，问这台包装机工作是否正常？

9. 玉米穗重服从正态分布，已知种在清洁区内玉米的平均穗重为 300 g，随机抽取污灌区内 7 个玉米穗重(单位：g)，分别为 298，290，297，301，299，297，292。问污灌对玉米穗重是否有显著影响？

10. 某城市在不同季节、不同地点采样分析大气中飘尘含量，结果见表 1.33，试分析大气中飘尘含量的时空差异是否显著。

表 1.33　　某市大气飘尘监测结果　单位：mg/m³

	春季	夏季	秋季	冬季
市中区	0.620	0.420	0.880	1.20
近郊区	0.614	0.475	0.667	1.15
远郊区	0.379	0.200	0.540	0.94

【参考文献】

[1] 盛聚，谢式千，潘承毅．概率论与数理统计［M］．北京：高等教育出版社，2004.
[2] 何晓群．现代统计分析方法与应用［M］．北京：中国人民大学出版社，2003.
[3] 于义良，张银生．实用概率统计［M］．北京：中国人民大学出版社，2002.
[4] 陈玉成，吕宗清，李章平．环境数学分析［M］．重庆：西南师范大学出版社，1998.
[5] 杨虎，刘琼荪，钟波．数理统计［M］．北京：高等教育出版社，2006.
[6] 康永尚，沈金松，谌卓恒．现代数学地质［M］．北京：石油工业出版社，2005.
[7] 贾俊平．统计学［M］．北京：清华大学出版社，2005.

第2章　环境一元线性回归分析

任何科学研究都要揭示客观世界内在的本质联系，除了要研究定性关系，还应尽可能建立定量关系，这种定量关系常用模型(函数)形式表现。例如，因变量 y 是某点 SO_2 的浓度，自变量 x 是排放源的排放量，通过建立定量分析模型，如 $y=f(x)$，即可分析自变量作用的大小，求得什么情况下污染严重，什么情况下污染较轻。如果给出未来排放源的排放量，通过模型还可以预测未来该点 SO_2 的浓度，并据此制定防治措施。环境科学研究中，相当一部分是对环境问题进行分析、预测。一元线性回归是描述2个变量之间统计关系的一种最简单的统计分析技术。通过建立一元线性回归模型，我们可以很好地了解回归分析的统计思想并解决实际环境问题。

本章的主要内容是：

- 一元线性回归的建模原理；
- 模型参数的最小二乘估计；
- 线性回归方程的显著性检验；
- 线性回归式的误差估计；
- 可化为一元线性回归的曲线回归；
- 环境应用。

2.1　一元线性回归模型

2.1.1　变量间的统计关系

实际生活中常常会遇到多个变量在同一个过程之中，它们相互联系、相互制约的情形。有的变量间存在完全确定的函数关系，例如圆面积与半径之间有确定的关系式。还有一些变量间存在不完全确定的关系，例如正常人的血压和年龄之间有一定关系。污染物排放浓度与温度大致成直线关系，但不能精确地表示出来。其实，它们是随机变量(或至少其中一个是随机变量)之间的关系，常称为统计关系或相关关系。为了深入了解事物的本质，往往需要寻找这些变量间的依存关系式。

现象间的依存关系大致可分为两种类型：函数关系和统计关系（何晓群，2003）。

（1）函数关系。函数关系是指现象之间一种严格的、确定性的依存关系，表现为某一现象发生变化，另一现象也随之发生变化，且有确定的值与之相对应。例如，银行的 1 年期存款利率为年息 1.98%，存入的本金为 x，到期本息为 y，则 $y=x+1.98\%x$（不考虑利息税）；再如，某种股票的成交额 y 与该股票的成交量 x、成交价格 p 之间的关系可以用 $y=px$ 来表示，这都是函数关系。

（2）统计关系。统计关系是指客观现象之间确实存在，但数量上不是严格对应的依存关系，表现为某一现象的每一数值，可以有另一现象的若干数值与之相对应。例如，成本的高低与利润的多少有密切关系，但某一确定的成本与相对应的利润却是不确定的。因为影响利润的因素除成本外，还有价格、供求平衡、消费嗜好以及其他偶然因素等的影响。

函数关系和统计关系既有区别，又有联系。有些函数关系因为观察或测量误差以及各种随机因素的干扰等，常常通过统计关系表现出来；而在统计关系中，对其数量间的规律性了解得越深刻，统计关系越有可能转化为函数关系或借助函数关系来表现。

统计关系规律性的研究是统计学研究中的主要对象，目前关于统计关系的研究已形成统计学中两个重要的分支，即相关分析和回归分析。

相关分析和回归分析是研究事物的相互关系、测定它们之间联系的紧密程度、揭示其变化的具体形式和规律性的统计方法，是构造各种环境模型、进行结构分析、政策评价、预测和控制的重要工具。通过相关分析，可以判断两个或两个以上的变量之间是否存在相关关系、相关关系的方向、形态及相关关系的密切程度。回归分析是对具有相关关系现象间数量变化的规律性进行测定，确立一个回归方程，并对所建立的回归方程的有效性进行分析和判断，以便进一步进行估计和预测。两者之间既有联系又有区别。

（1）相关分析和回归分析之间的联系

①理论和方法具有一致性。

②无相关就无回归，相关程度越高，回归越好。

③相关系数和回归系数方向一致，可以互相推算。

（2）相关分析和回归分析之间的区别

①相关分析中 x 与 y 对等，回归分析中 x 与 y 要确定自变量和因变量。

②相关分析中 x，y 均为随机变量，回归分析中只有 y 为随机变量。

③相关分析测定相关程度和方向，回归分析用回归模型进行预测和控制。

由于相关分析和回归分析的研究侧重不同，使得它们的研究方法也大不一

样。回归分析已成为环境统计分析中应用最活跃的分支之一。

2.1.2 一元线性回归模型

"回归"一词是由英国生物学家 F. Galton(1822—1911)在研究人体身高的遗传问题时首先提出的。根据遗传学的观点，子辈的身高受父辈影响，以 x 记父辈身高，y 记子辈身高。虽然子辈身高一般受父辈影响，但同样身高的父亲，其子身高并不一致。因此，x 和 y 之间存在一种统计关系。一般而言，父辈身高者，其子辈身高也高。依此推论，祖祖辈辈遗传下来，身高必然向两极分化，而事实上并非如此，显然有一种力量将身高拉向中心，即子辈的身高有向中心回归的特点，"回归"一词即源于此。虽然这种向中心回归的现象只是特定领域里的结论，并不具有普遍性，但从它所描述的关于 x 为自变量，y 为不确定的因变量这种变量间的关系看，和我们现在的回归含义是相同的。不过现代回归分析虽然沿用了"回归"一词，但内容已有很大变化，它是一种应用于许多领域的、广泛的统计分析方法，在环境科学理论和实验研究中也发挥着重要的作用。

回归分析通过一个变量或一些变量的变化来解释另一变量的变化。其主要内容和步骤是：首先，根据对问题的分析判断，将变量分为自变量和因变量；其次，设法找出合适的数学方程式(即回归模型)描述变量间的关系；由于涉及的变量具有不确定性，接着还要对回归模型进行统计检验；最后，利用回归模型，根据自变量去估计、预测因变量。

回归有不同种类，按照自变量的个数分，有一元回归和多元回归。只有一个自变量的叫一元回归，有两个或两个以上自变量的叫多元回归；按照回归曲线的形态分，有线性(直线)回归和非线性(曲线)回归。实际分析时应根据客观现象的性质、特点、研究目的和任务选取回归分析的方法。本节仅讨论一元线性回归模型。

例 2.1 某河流溶解氧浓度(以百万分之一计)随着流动时间而下降，现测得 8 组数据，如表 2.1 所示。

表 2.1 河流中溶解氧浓度

流动时间 x/d	0.5	1.0	1.6	1.8	2.6	3.2	3.8	4.7
溶解氧浓度 y	0.28	0.29	0.29	0.18	0.17	0.18	0.10	0.12

由表 2.1 所示，首先根据表中提供的数据，画出以流动时间为横坐标，溶解

氧浓度为纵坐标的散点图(图 2-1)，从散点图上可以很明显地观察出各个点都近似均匀地分布在一条直线的周围，但是又不完全在一条直线上。

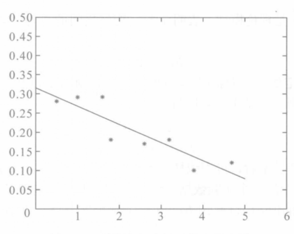


图 2-1　溶解氧浓度随时间变化曲线

如果流动时间和溶解氧浓度之间是线性相关关系，则引起这些点(x_i, y_i)与直线偏离的主要原因是实际测量过程中存在的不可控因素。它们影响了实验数据，产生了一定的误差。流动时间的实验数据和溶解氧浓度的实验数据可用线性关系式表示。

对于一个实际问题，通常先将实测数据在直角坐标系内描述成散点图。如果实测点基本上在一条直线附近波动，则自变量和因变量之间可以用线性关系描述，就可以采用线性模型。经判断，具有线性关系的两个变量 y 与 x，可构造一元线性回归模型为：

$$y=a+bx+\varepsilon$$

其中，a 与 b 为模型参数，x 是自变量，y 是因变量，ε 为随机误差项，b 称为回归系数。

假定 $E(\varepsilon)=0$，有回归函数：

$$\mu(x)=E(y)=a+bx \tag{2.1}$$

其中，截距 a 表示在没有自变量 x 的影响时，其他各种因素对因变量 y 的平均影响；回归系数 b 表明自变量 x 每变动一个单位，因变量 y 平均变动 b 个单位。

2.1.3 最小二乘法估计

$\mu(x)=a+bx$ 是理论模型，表明 x 与 y 变量之间的平均变动关系，而变量 y 的实际值应为：

$$y_i=(a+bx_i)+\varepsilon_i=\mu(x_i)+\varepsilon_i$$

其中，a，b 的确定如下：为获得 a，b 的估计，需对自变量及与其对应的因变量进行 n 次独立观测。假设实测数据为$(x_i, y_i)(i=1, 2, \cdots, n)$，即：

$$\begin{matrix} x_1, & x_2, & x_3, & \cdots, & x_n \\ y_1, & y_2, & y_3, & \cdots, & y_n \end{matrix} \tag{2.2}$$

确定采用线性模型后，可用线性关系式 $\mu(x)=a+bx$ 对实验数据进行拟合。如何根据式(2.2)中的 n 个实验数据来估计 a，b? 自然会想到，要选择 a，b，使得 x_i 直线上对应的值 $\mu(x_i)=a+bx_i$ 与 x_i 对应的实测值 y_i 的误差 ε_i 在某种意义下最小。

$$\varepsilon_i=y_i-\mu(x_i) \qquad (i=1, 2, \cdots, n)$$

上式也可以写为：

$$y_i=a+bx_i+\varepsilon_i \qquad (i=1, 2, \cdots, n) \tag{2.3}$$

ε_i 称为第 i 次观测的随机误差，设 ε_i 相互独立，即 $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)=0$，$(i, j=1, 2, \cdots, n)$且期望 $E(\varepsilon_i)=0$，$D(\varepsilon_i)=\sigma^2$，$\mathrm{Cov}(\varepsilon_i, x_i)=0$ $(i=1, 2, \cdots, n)$，所以 $\varepsilon\sim N(0, \sigma^2)$。这就意味着 y 的数学期望是 x 的线性函数，而且 $y\sim N(a+bx, \sigma^2)$。

显然 ε_i 越小，方程对数据拟合得越理想，但 ε_i 有正负之分，为避免正负抵消，可求 $\sum_{i=1}^{n}[y_i-\mu(x_i)]^2=\sum_{i=1}^{n}\varepsilon_i^2$ 的最小值。下面采用最小二乘法原则来估计 a，b。

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2 \tag{2.4}$$

利用极值法使 Q 达到最小值，于是问题化为解方程组：

$$\begin{cases} \dfrac{\partial Q}{\partial a}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]=0 \\ \dfrac{\partial Q}{\partial b}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]x_i=0 \end{cases} \tag{2.5}$$

此方程组称为正规方程组，由此方程组解得 a，b 的最小二乘估计值 $\hat{a}$，$\hat{b}$。

解正规方程组可得：

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\qquad \hat{a}=\bar{y}-\hat{b}\bar{x} \tag{2.6}$$

其中，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，这里 $\bar{x}$，$\bar{y}$ 为观测变量的均值。

令 $L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$，$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$，$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$

则

$$\hat{b}=\frac{L_{xy}}{L_{xx}},\ \hat{a}=\bar{y}-\hat{b}\bar{x} \tag{2.7}$$

取 $\hat{a}+\hat{b}x$ 作为 $\mu(x)=a+bx$ 的估计，记 $\hat{y}=\hat{a}+\hat{b}x$。

这样所得方程 $\hat{y}=\hat{a}+\hat{b}x$ 称为经验回归方程，简称回归方程。由于参数的估计结果是通过最小二乘法得到的，故称为普通最小二乘估计量(ordinary least squares estimators，OLSE)。

根据公式，求例 2.1，可得：

$$\bar{x}=\frac{1}{8}\sum_{i=1}^{8}x_i=2.4,\ \bar{y}=\frac{1}{8}\sum_{i=1}^{8}y_i=1.61\times\frac{1}{8}=0.20$$

$$L_{xx}=14.5000,\ L_{xy}=-0.6840,\ L_{yy}=0.0407$$

由此可得：

$$\hat{b}=\frac{L_{xy}}{L_{xx}}=-0.0472,\ \hat{a}=\bar{y}-\hat{b}\bar{x}=0.3145$$

这样就得到流动时间(x)和溶解氧浓度(y)之间的线性关系式：

$$\hat{y}=\hat{a}+\hat{b}x=0.3145-0.0472x$$

2.2　线性回归方程的显著性检验

在很多实际研究过程中，我们事先并不知道变量之间是否存在线性关系，这时可以采用第一节所介绍的方法，先将实验数据描绘成散点图，再根据观测点是否基本上都在一条直线附近来判断变量之间是否为线性关系，最后利用最小二乘估计法估计出 $\hat{a}$ 和 $\hat{b}$，即可得到经验回归方程。为了判断上述估计法得到的经验回归方程是否能够精确地描述两个变量之间的关系，本节将采用假设检验的方法进行统计推断。

2.2.1 F 检验法

我们知道 y_i(实测值)与 $\hat{y}_i$(回归值或计算值)之间之所以有差异，一般是由下述两个原因引起的：一是当 y 和 x 之间的确有线性关系时，由实验过程中的随机误差引起；二是当 y 和 x 之间不存在线性关系而引起的 y_i 与 $\hat{y}_i$ 之间的不同。

由于各种原因引起因变量的总波动，称之为总偏差平方和，用 $S_{总}$ 表示。

$$\begin{aligned} S_{总} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ &= S_{残}+S_{回} \end{aligned}$$

其中，$S_{残}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，称为残差平方和（或剩余平方和）；$S_{回}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，称为回归平方和；$\hat{y}_i=\hat{a}+\hat{b}x_i$。

交叉乘积项：
$$\begin{aligned} 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= 2\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i-2\sum_{i=1}^{n}(y_i-\hat{y}_i)\bar{y} \\ &=2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{a}+\hat{b}x_i)-2\sum_{i=1}^{n}(y_i-\hat{y}_i)\bar{y} \\ &=2\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{a}+2\hat{b}\sum_{i=1}^{n}(y_i-\hat{y}_i)x_i-2\sum_{i=1}^{n}(y_i-\hat{y}_i)\bar{y} \end{aligned}$$

由正规方程组可知：$\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{a}=0$；$\sum_{i=1}^{n}(y_i-\hat{y}_i)x_i=0$；$\sum_{i=1}^{n}(y_i-\hat{y}_i)\bar{y}=0$

所以可以得到：$2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$

这样可得：$S_{总}=S_{残}+S_{回}$

$S_{残}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 是 y 的实际值与回归值之差的平方和，它是由随机因素以及测量误差引起的，它的大小反映了测量误差及其随机因素对 y 的影响，是总偏差平方和中不能被回归方程解释的部分。

$S_{回}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 是 y 的回归值与平均值之差的平方和，它反映了自变量

x_1，x_2，…，x_k 的变化所引起的 y 的波动。$S_回$ 是总偏差平方和中由回归方程解释的部分。

如果变量 x 和 y 之间无线性关系，则 $b=0$，这相当于检验假设：H_0：$b=0$ 是否成立。

在一般的线性模型中，当假设 H_0 为真时，一切 $y_i \sim N(a, \sigma^2)$，并且相互独立，由此容易得到：

$$\bar{y} \sim N\left(a, \frac{1}{n}\sigma^2\right); \; E(y_i^2)=D(y_i)-E^2(y_i)=\sigma^2-a^2;$$

$$E(\bar{y}^2)=D(\bar{y})-E^2(\bar{y})=\frac{1}{n}\sigma^2-a^2$$

所以

$$\begin{aligned} E(S_总) &= E\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\ &= E\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \\ &= \sum_{i=1}^{n} E(y_i^2) - nE(\bar{y}^2) \\ &= n(\sigma^2-a^2)-n\left(\frac{1}{n}\sigma^2-a^2\right)=(n-1)\sigma^2 \end{aligned}$$

因此，$\frac{S_总}{n-1}$ 是 σ^2 的无偏估计量。

容易证明，在 H_0 为真时，

$$\frac{S_总}{\sigma^2} \sim \chi^2(n-1); \; \frac{S_残}{\sigma^2} \sim \chi^2(n-2); \; \frac{S_回}{\sigma^2} \sim \chi^2(1)$$

在 $S_残$，$S_回$ 相互独立的条件下，根据 F 分布的定义可知：

$$F=\frac{S_回}{S_残/(n-2)} \sim F(1, n-2)$$

这就是用来检验假设 H_0 的统计量，按照一般显著性检验的程序，在给定显著性水平 α 的前提下，查 $F(1, n-2)$ 分布表，可得临界值 $F_\alpha(1, n-2)$。

(1) 当 $\frac{S_回}{S_残/(n-2)} \geq F_\alpha(1, n-2)$，则否定 H_0：$b=0$，认为一元线性回归式可用，即变量 x 和 y 之间的关系可以用线性关系代替。

(2) 当 $\frac{S_回}{S_残/(n-2)} < F_\alpha(1, n-2)$，则接受 H_0：$b=0$，认为一元线性回归式无意义，即变量 x 和 y 之间的关系不能用线性关系来代替。

根据例 2.1，可求得：

$S_总=L_{yy}=0.040\,7$，$S_回=0.032\,3$，$S_残=S_总-S_回=0.040\,7-0.032\,3=0.008\,4$

则：$F=\frac{S_{回}}{S_{残}/(n-2)}=\frac{0.032\ 3}{0.008\ 4/6}=23.071\ 4$

当 $\alpha=0.05$ 时，查 F 分布表，得到 $F_{\alpha}(1, n-2)=5.990\ 0$，由于 $F=23.071\ 4>5.990\ 0=F_{\alpha}(1, n-2)$，所以 $\hat{y}=\hat{a}+\hat{b}x=0.314\ 5-0.047\ 2x$ 线性关系显著。

2.2.2 相关系数检验法

上面介绍了用于显著性检验的 F 检验法，下面我们将介绍另一种用于显著性检验的方法——相关系数法。

因为

$$
\begin{aligned}
S_{残} &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)]^2 \\
&= \sum_{i=1}^{n}[y_i-(\bar{y}-\hat{b}\bar{x}+\hat{b}x_i)]^2 \\
&= \sum_{i=1}^{n}[(y_i-\bar{y})-\hat{b}(x_i-\bar{x})]^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2-2\hat{b}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2-2\hat{b}\sum_{i=1}^{n}\hat{b}(x_i-\bar{x})^2+\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2=L_{yy}-\hat{b}^2L_{xx}
\end{aligned}
$$

显然 $S_{残}\geqslant 0$。当 $S_{残}=0$ 时，说明该直线与实际情况完全吻合，y 与 x 之间显然是线性关系；但如果出现 $S_{残}\neq 0$，$\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2=0$，这时 $S_{残}=\sum_{i=1}^{n}(y_i-\bar{y})^2$，说明 $S_{残}$ 的变化与 x 无关，从而 y 是平行于 x 轴的直线，因而 x 和 y 是处于零相关，也即 x 和 y 是不相关的。在实际衡量 y 与 x 之间的相关性时，由于 $S_{残}$ 与 y^2 是同一量纲，有时数值差别很大也难以说明拟合的密切程度，因此我们必须寻找统一的衡量标准，消除量纲，把数值标准化。由上述过程的启发，我们可以令：

$$
r^2=\frac{\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-S_{残}}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{S_{残}}{L_{yy}}
$$

因为 $L_{yy}\geqslant S_{残}$，所以 $r\leqslant 1$。当 $r=1$ 时，为完全相关，$r=0$ 时为零相关。这样很容易得到，如果 y 与 x 的相关性较好，则 $S_{残}$ 值较小，从而 $r^2\to 1$；如果 x

与 y 相关性较差，则 $S_{残}$ 值较大，从而 $r^2 \to 0$。

因为 $\hat{b}=\frac{L_{xy}}{L_{xx}}$，所以有：

$$r^2=\frac{\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\hat{b}^2L_{xx}}{L_{yy}}=\frac{L_{xy}^2}{L_{xx}^2}\cdot\frac{L_{xx}}{L_{yy}}=\frac{L_{xy}^2}{L_{xx}L_{yy}},r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

因为 L_{xx}，$L_{yy}>0$，所以 $\hat{b}$，r 均与 L_{xy} 同符号。这样，我们找到了一个无量纲 r 且能反映出相关程度的相关系数。当 $r<0$ 时，称为负相关；当 $r>0$ 时，称为正相关。注意，r 仅仅反映 x，y 的样本间的线性相关程度，它是个统计量。若要对总体相关系数进行统计推断，还需要进行假设检验。

究竟 r 多大时，才可以认为 y 与 x 之间具有显著的线性相关关系呢？这需要将 r 与其临界值作比较，利用相关系数检验法对线性回归分析进行显著性检验，对于给定的显著性水平 α，由样本计算得到 $r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$。

(1)若 $|r|\geqslant r_{\alpha}(n-2)$，则认为线性回归是显著的，$y$ 与 x 之间可以认为是线性相关关系。

(2)若 $|r|<r_{\alpha}(n-2)$，则认为线性回归不显著，y 与 x 之间不存在线性相关关系。

根据例 2.1，可以得到：$L_{xx}=14.500\ 0$，$L_{xy}=-0.684\ 0$，$L_{yy}=0.040\ 7$

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{-0.684\ 0}{\sqrt{14.500\ 0\times 0.040\ 7}}=-0.890\ 4$$

取显著性水平 $\alpha=0.05$，按照自由度 $n-2=6$ 查相关系数表，得 $r_{\alpha}(n-2)=0.707$。由于 $|r|=0.890\ 4>0.707$，故认为 y 与 x 之间的线性关系较显著，即 $\hat{y}=\hat{a}+\hat{b}x=0.314\ 5-0.047\ 2x$ 可以表达 y 与 x 之间存在的线性相关关系。显然这一检验结果与 F 检验法的结果一致。

2.2.3　样本决定系数 r^2

由回归平方和与残差平方和的意义我们知道，在总偏差平方和中，如果回归平方和所占的比重越大，则线性回归效果越好；如果残差平方和所占的比重大，则回归直线与样本观测值拟合的就不理想。这里把回归平方和与总偏差平方和之比定义为样本决定系数，记作 r^2。r^2 是一个回归直线与样本观测值拟合优度判定的指标，r^2 的值总在 0 和 1 之间。一个线性回归模型如果充分利用了 x 的信息，

则 r^2 越大，拟合优度就越好；反之，若 r^2 不大，说明模型中给出的 x 对 y 的信息还不够充分，应进行修改，使 x 对 y 的信息得到充分利用。例如，决定系数为：$r^2=0.9711$，这说明在 y 值与 $\bar{y}$ 的偏差的平方和中有 97.11%可以通过变量 x 来解释。

在一元线性回归中，容易证明 F 检验法与相关系数检验法其实质是相同的，两者检验的结果也是一致的。因此，在线性回归的显著性检验中，可以选择 F 检验法或相关系数检验法。

2.3 线性回归式的误差估计

2.3.1 线性回归式的误差估计

由经验知，y 与 x 之间的线性关系显著时，则可以认为回归方程 $\hat{y}=\hat{a}+\hat{b}x$ 反映了 y 与 x 之间的变化规律，这时可以利用回归方程对 y 进行误差估计。

当 $x=x_0$ 时，相应的 y_0 是一个随机变量，利用回归方程对 y_0 作预测，$\hat{y}_0=\hat{a}+\hat{b}x_0$ 就是 y_0 的一个预测值，这种预测值称为点预测。$y_0-\hat{y}_0$ 用来表示实际值和估计值之间的误差。

设 $\hat{y}=\hat{a}+\hat{b}x$ 是由样本$(x_i,\ y_i)(i=1,\ 2,\ \cdots,\ n)$根据式(2.1)按照最小二乘法估计的线性回归方程，$\hat{y}_0=\hat{a}+\hat{b}x_0$ 是当 $x=x_0$ 时相应的变量 y_0 的估计量，则容易证明：

$$\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}}\sim t(n-2)$$

由 t 分布的分位点概念，对于给定的显著性水平 α，则：

$$P\left\{\left|\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}}\right|\leqslant t_\alpha(n-2)\right\}=1-\alpha$$

其中，总体方差 σ^2 的一个无偏估计量 $\hat{\sigma}^2$ 可以通过如下方式求得：

$$\hat{\sigma}^2=\frac{S_{残}}{n-2}=\frac{L_{yy}-\hat{b}^2L_{xx}}{n-2}$$

由于 $r=\dfrac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$，$\hat{b}=\dfrac{L_{xy}}{L_{xx}}$

故 $\hat{\sigma}^2=\frac{L_{yy}-\hat{b}^2L_{xx}}{n-2}=\frac{(1-r^2)L_{yy}}{n-2}$

当 n 较大时，近似有(卢崇飞等，1988)：

$$y_0-\hat{y}_0\sim N(0,\ \hat{\sigma}^2)$$

于是，近似有：

$$0.95\approx P(\hat{y}_0-2\hat{\sigma}<y_0<\hat{y}_0+2\hat{\sigma})$$
$$0.99\approx P(\hat{y}_0-3\hat{\sigma}<y_0<\hat{y}_0+3\hat{\sigma})$$

通常，由回归方程计算的估计值能够满足 $0.95\approx P(\hat{y}_0-2\hat{\sigma}<y_0<\hat{y}_0+2\hat{\sigma})$，则误差可以忽略。

2.3.2　线性回归的步骤

通过以上的介绍和具体分析，现在将一元线性回归分析的主要步骤作如下总结：

(1)设变量 x 和 y 的线性回归方程为：$\hat{y}=\hat{a}+\hat{b}x$。

(2)求回归系数的估计量 $\hat{a}$，$\hat{b}$：$\hat{b}=\frac{L_{xy}}{L_{xx}}$，$\hat{a}=\bar{y}-\hat{b}\bar{x}$。

其中，均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，$x$ 偏差平方和 $L_{xx}=\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n$；$y$ 偏差平方和 $L_{yy}=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2/n$；$x$，$y$ 乘积的偏差和 $L_{xy}=\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)/n$。

(3)检验回归系数 $\hat{b}$ 是否为零。

检验 H_0：$b=0$ 成立时，令：

$$F=\frac{S_{回}}{S_{残}/(n-2)}\sim F(1,\ n-2)$$

给定显著性水平 α，查 $F(1,\ n-2)$分布表，可得临界值 $F_\alpha(1,\ n-2)$。

当$\frac{S_{回}}{S_{残}/(n-2)}\geqslant F_\alpha(1,\ n-2)$，则否定 H_0：$b=0$，认为一元线性回归式可用；当$\frac{S_{回}}{S_{残}/(n-2)}<F_\alpha(1,\ n-2)$，则接受 H_0：$b=0$，认为一元线性回归式无意义。

(4)求相关系数并作相关性检验。

首先求得相关系数 $r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$。

检验方法如下：

若 $|r| \geqslant r_{\alpha}(n-2)$，则线性回归是显著的，$y$ 与 x 之间是线性关系；

若 $|r| < r_{\alpha}(n-2)$，则认为线性回归不显著，y 与 x 之间不存在线性相关关系。

2.4 可化为一元线性回归的曲线回归

大量的环境监测数据表明，很多环境参数间不成正比例变化，即不具有线性相关性，这可以通过对它们作线性相关性检验时证明，这时应当按照它们之间具有的曲线关系处理监测数据(张孟威等，1989)。如果因变量 y 与自变量 x 之间关系的散点图已经明显地呈现非线性关系，或者是线性回归假设检验后发现它们不是线性关系，又或者是从专业的角度判断它们不可能是线性关系，但 y 与未知参数 a，b 之间的关系都是线性的。注意，线性回归是针对参数而言，而不是针对自变量而言。因此，有些因变量 y 对自变量 x 的曲线关系情形可以通过变量代换转换成线性的形式。

具体思路是通过作散点图或定性分析认为两个变量之间存在的相关关系为曲线相关时，可先根据变量间不同类型配合一条与其相适应的回归曲线，如指数曲线、双曲线等，然后再确定回归方程中的未知参数。对于那些可线性化的回归方程，对新变量而言，线性化后的方程都为直线方程，故其参数的确定可用线性回归方程求参数的公式计算。

下面就列举几个常用的转化方法。

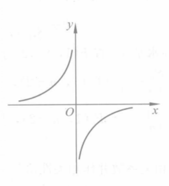
2.4.1 倒数变换

(1)双曲线(图 2-2)

y

O x

($k>0$)



($k<0$)

图 2-2 双曲线示意图

反比例函数　　$\frac{1}{y}=\frac{k}{x}$　（$k\neq0$，$x\neq0$，$y\neq0$）　　(2.8)

令　$y'=\frac{1}{y}$，$x'=\frac{1}{x}$；则得

$$y'=kx'$$

(2)S 型曲线(图 2－3)

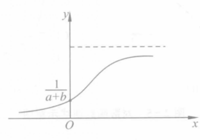


图 2－3　S 型曲线示意图

$$y=\frac{1}{a+be^{-x}}\qquad (a+be^{-x}\neq0) \tag{2.9}$$

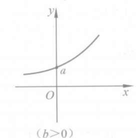
令　$y'=\frac{1}{y}$，$x'=e^{-x}$，则得

$$y'=a+bx'$$

2.4.2　对数变换

一般情况下，对于指数曲线、对数曲线、幂函数曲线，都用对数变换的方法将曲线回归分析转化为线性回归分析。

(1)指数函数(图 2－4)



y　a　O　x　($b<0$)

图 2－4　指数函数曲线示意图

$$y=ae^{bx}\quad (a>0) \tag{2.10}$$

对其两边取自然对数，得 $\ln y=\ln a+bx$

令　$y'=\ln y$，则得 $y'=\ln a+bx$

(2)对数函数(图 2-5)

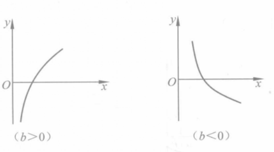


图 2-5　对数函数曲线示意图

$$y=a+b\lg x\quad (x>0) \tag{2.11}$$

令 $x'=\lg x$，则得 $y=a+bx'$

(3)幂函数(图 2-6)

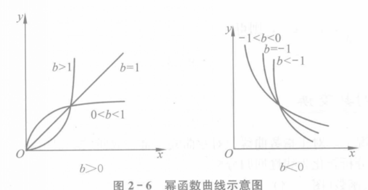


图 2-6　幂函数曲线示意图

$$y=ax^b\quad (a>0,\ x>0) \tag{2.12}$$

对上式两边取对数，得 $\lg y=\lg a+b\lg x$

令　$y'=\lg y$，$x'=\lg x$，则得 $y'=\lg a+bx'$

2.4.3　混合变换

有的方程无法用单一的变换线性化，需要用多种变换来实现。例如函数 $1/y=ce^{bx}(c>0)$ 的线性化过程就要涉及多种变换。首先令 $y_1=1/y$，得到 $y_1=$

ce^{bx}；再令 $y_2=\ln y_1$，得到 $y_2=a+bx$，其中 $a=\ln c$。这样就通过两种变换实现了函数 $1/y=ce^{bx}$ 的线性化。

2.5 环境应用

下面介绍几个利用一元线性回归分析来解决某些实际环境应用过程中经常遇到的问题。

例 2.2 环境气象数据的回归计算

大气环境污染问题中，风速往往是一个重要因素。在城镇气象数据中，常可从当地气象站了解到日平均风速，而城镇街道距地面 2 m 高空中的风速，可由气象站测得的风速推算得到。推算的方法是，预先测定 n 对风速数据，每对风速由日平均风速与街道风速组成（同步数据），然后求得这两种风速间的回归方程式，由回归方程式便可推算得街道风速(张孟威等，1989)。

例如某城镇实测的风速数据如表 2.2 所示，试计算两种风速间的线性方程式，并作线性相关性试验，以及计算残差平方和。

表 2.2　　**风速监测数据　单位：m/s**

实验编号	1	2	3	4	5	6	7	8	9	10
日平均风速 x	4.3	2.7	3.3	4.7	4.3	5.7	6.0	6.0	5.3	6.0
街道风速 y	3.0	3.5	3.5	4.0	4.5	2.5	4.0	4.0	4.5	4.5
实验编号	11	12	13	14	15	16	17	18	19	20
日平均风速 x	5.0	3.7	2.7	1.0	1.0	2.7	1.0	0.7	0.7	0.7
街道风速 y	3.5	2.2	1.8	1.2	1.0	2.0	1.2	1.0	1.0	0.4

解　(1)先作散点图（图 2－7）。

由图 2－7 可知，所有的点基本上都分布在一条直线周围，故可以采用线性回归分析。

(2)根据表 2.2 中的数据，可以得到：

$$\bar{x}=\frac{1}{20}\sum_{i=1}^{20}x_i=3.375\ 0,\ \bar{y}=\frac{1}{20}\sum_{i=1}^{20}y_i=2.665\ 0$$

$$\sum_{i=1}^{20}x_i^2=303.570\ 0,\ \sum_{i=1}^{20}y_i^2=178.870\ 0,\ \sum_{i=1}^{20}x_iy_i=226.130\ 0$$

$$L_{xy}=\sum_{i=1}^{20}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{20}x_iy_i-20\bar{x}\bar{y}=46.242\ 5$$

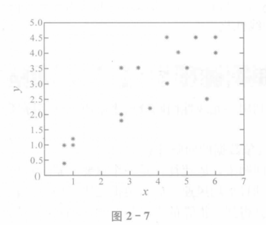


图 2-7

$$L_{xx}=\sum_{i=1}^{20}(x_i-\bar{x})^2=\sum_{i=1}^{20}x_i^2-20\bar{x}^2=75.757\ 5$$

$$L_{yy}=\sum_{i=1}^{20}(y_i-\bar{y})^2=\sum_{i=1}^{20}y_i^2-20\bar{y}^2=36.825\ 5$$

$$\hat{b}=\frac{L_{xy}}{L_{xx}}=\frac{46.242\ 5}{75.757\ 5}=0.610\ 4$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}=2.665\ 0-0.610\ 4\times3.375\ 0=0.604\ 9$$

所以，得到的线性回归方程为：$\hat{y}=\hat{a}+\hat{b}x=0.604\ 9+0.610\ 4x$

(3)$S_{总}=L_{yy}=36.825\ 5$，$S_{回}=28.226\ 3$，$S_{残}=S_{总}-S_{回}=8.599\ 2$

则 $F=\dfrac{S_{回}}{S_{残}/(n-2)}=\dfrac{28.226\ 3}{8.599\ 2/18}=59.084\ 2$

当 $\alpha=0.05$ 时，查 F 分布表，得到 $F_\alpha(1,\ n-2)=4.41$，由于 $F=59.084\ 2>4.41=F_\alpha(1,\ n-2)$，所以认为线性关系式 $\hat{y}=\hat{a}+\hat{b}x=0.604\ 9+0.610\ 4x$ 显著。

因为：$L_{xx}=75.757\ 5$，$L_{xy}=46.242\ 5$，$L_{yy}=36.825\ 5$

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{46.242\ 5}{\sqrt{75.757\ 5\times36.825\ 5}}=0.875\ 5$$

取显著性水平 $\alpha=0.05$，按照自由度 $n-2=18$ 查相关系数检验表（附表 2），得 $r_\alpha(n-2)=0.444$。由于 $|r|=0.875\ 5>0.444$，故认为 y 与 x 之间的线性关系较显著，即 $\hat{y}=\hat{a}+\hat{b}x=0.604\ 9+0.610\ 4x$ 可以表达 y 与 x 之间存在的线性相关关系。显然这一检验结果与 F 检验法的结果一致。

例 2.3 一个非线性回归的例子

絮凝体沉降随时间(x)的去除率(y)符合指数函数 $y=ae^{b/x}\cdot\varepsilon(\ln\varepsilon\sim N(0,\sigma^2))$。现得实验数据如表 2.3(陈玉成等，1998)，据此建立指数回归方程，并进行检验。

表 2.3 絮凝体沉降时的去除率

x/min	5	10	15	20	25	30	60
y/%	38	51	58	62	64	65	67

解 (1)根据表 2.3 的数据画出散点图（图 2-8）。

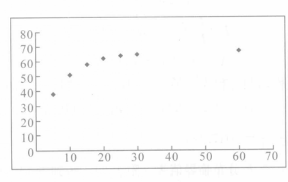


图 2-8 表 2.3 数据的散点图

(2)作变换。

对 $y=ae^{b/x}\cdot\varepsilon$ 两边取自然对数，得 $\ln y=\ln a+b\cdot\dfrac{1}{x}+\ln\varepsilon$

令 $y'=\ln y$，$x'=\dfrac{1}{x}$，则得 $y'=A+Bx'+\varepsilon'(\varepsilon'\sim N(0,\sigma^2))$

其中，$A=\ln a$，$B=b$

记 $\hat{y}'=\hat{A}+\hat{B}x'$

将变换后的数据画出散点图（图 2-9）。

由图 2-9 可知，所有的点基本上都分布在一条直线周围，故可以采用线性回归分析。

(3)根据变换后的数据，采用一元线性回归可以得到：

$$\overline{x}'=0.072\,4,\ \overline{y}'=4.042\,1$$

$$L_{x'y'}=-0.074\,7$$

$$L_{x'x'}=0.023\,3$$

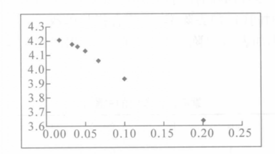


图 2-9　变换后数据的散点图

$$L_{y'y'}=0.240\ 9$$

$$\hat{B}=\frac{L_{x'y'}}{L_{x'x'}}=-3.210\ 8$$

$$\hat{A}=\bar{y}'-\hat{B}\bar{x}'=4.274\ 5$$

所以得到的线性回归方程为：$\hat{y}'=\hat{A}+\hat{B}x'=4.274\ 5-3.210\ 8x'$　　(2.13)

$$S_{回}=L_{yy}=0.239\ 8,\ S_{残}=0.001\ 1$$

则 $F=\frac{S_{回}}{S_{残}/(n-2)}=1\ 047.187\ 2$

当 $\alpha=0.05$ 时，查 F 分布临界值表（附表5），得到 $F_{\alpha}(1,\ n-2)=6.61$，由于 $F=1\ 047.187\ 2>6.61$，所以认为线性关系式 $\hat{y}'=\hat{A}+\hat{B}x'=4.274\ 5-3.210\ 8x'$ 显著。

$$|r|=\frac{|L_{x'y'}|}{\sqrt{L_{x'x'}L_{y'y'}}}=0.997\ 6$$

取显著性水平 $\alpha=0.05$，按照自由度 $n-2=5$ 查相关系数检验表（附表2），得 $r_{\alpha}(n-2)=0.754$。由于 $|r|=0.997\ 6>0.754$，故认为 y' 与 x' 之间的线性关系显著，即 $\hat{y}'=\hat{A}+\hat{B}x'=4.274\ 5-3.210\ 8x'$ 可以表达 y' 与 x' 之间存在的线性相关关系。显然这一检验结果与 F 检验法的结果一致。

将　$y'=\ln y$，$x'=\frac{1}{x}$ 代入式(2.13)得：

$$\hat{y}=71.846\ 6e^{-3.210\ 8/x}$$

【思考题 2】

1. 试述变量间统计关系和函数关系的本质区别。
2. 试述回归分析与相关分析的区别与联系。
3. 一元线性回归模型有哪些基本假定?
4. 一企业排水的 COD 及 BOD_5 的结果见表 2.4。

表 2.4　　COD 和 BOD_5 实测值

样品号	COD	BOD_5	样品号	COD	BOD_5
1	34.70	15.59	21	89.64	49.32
2	63.26	49.80	22	97.80	40.01
3	67.35	22.68	23	21.05	10.83
4	39.96	11.43	24	74.04	23.20
5	62.04	11.80	25	84.83	35.00
6	141.42	47.90	26	16.62	13.05
7	47.84	9.56	27	61.79	33.36
8	75.23	32.36	28	88.26	28.08
9	80.61	30.40	29	138.37	52.93
10	145.05	85.08	30	122.77	51.24
11	51.80	13.61	31	52.66	17.73
12	130.07	75.02	32	92.20	25.82
13	30.17	6.02	33	145.74	56.08
14	116.20	73.76	34	117.66	45.04
15	59.00	22.08	35	69.01	26.28
16	52.86	31.68	36	79.01	24.82
17	35.54	6.90	37	81.79	38.40
18	146.51	65.64	38	98.26	44.04
19	94.75	43.32	49	125.64	58.43
20	85.53	38.26	40	142.99	73.68

(1)画散点图；
(2)判断 COD 与 BOD_5 之间是否大致成线性关系；
(3)用最小二乘估计求回归方程；
(4)计算 COD 与 BOD_5 的决定系数；
(5)对回归方程作残差图，并作分析；
(6)计算当 COD=99 时，BOD_5 的值；
(7)给出置信水平为 95%的预测区间。

5. 在一项水分渗透实验中，得观测时间和水的重量的数据如表 2.5 所示。

表 2.5　观测时间和水的重量数据

观测时间 x/s	1	2	4	8	16	32	64
水的重量 y/g	4.22	4.02	3.85	3.59	3.44	3.02	2.59

(1)画出散点图；
(2)求曲线回归方程 $y=ax^b$；
(3)对 $\ln y$ 与 $\ln x$ 之间的线性回归关系进行显著性检验 $\alpha=0.05$。

6. 试用一元线性回归模型解决一个实际的环境问题。

【参考文献】

[1] 何晓群. 现代统计分析方法与应用 [M]. 北京：中国人民大学出版社，2003.

[2] 卢崇飞，高惠璇，叶文虎. 环境数理统计学应用及程序 [M]. 北京：高等教育出版社，1988.

[3] 张孟威，康德梦. 环境问题的数学解法及计算机应用 [M]. 北京：中国环境科学出版社，1989.

[4] 陈玉成，吕宗清，李章平. 环境数学分析 [M]. 重庆：西南师范大学出版社，1998.

第3章　环境多元线性回归分析

在环境科学中，经常要研究多个自变量对我们关心的指标值的影响，一般要根据已知数据来建立多个自变量与指标值之间的数量关系式，仅用一元线性回归分析方法是远远不够的。实际应用回归分析法时，常需要有更一般的模型，把两个或更多个解释变量的影响分别估计在内，即多元回归。当影响因素与因变量之间是线性关系时，所进行的多元回归分析就是多元线性回归，即多元线性回归是研究一个因变量和多个自变量之间数量上相互依存的线性关系。

本章的主要内容是：

- 多元线性回归模型；
- 参数的最小二乘估计；
- 回归方程的显著性检验；
- 回归系数的显著性检验；
- Matlab 语言在多元线性回归中的应用；
- 环境应用。

3.1　多元线性回归模型

多元线性回归分析除了计算比一元线性回归分析复杂外，其他的都同一元线性回归类似。因此，我们可以假设随机变量 y 与自变量 x_1，x_2，…，x_k 之间有如下的线性关系：

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k+\varepsilon \tag{3.1}$$

其中，k 为自变量的个数，ε 为随机误差项，b_0，b_1，…，b_k 称为回归系数。对于式(3.1)有如下假设：

(1)自变量 x_i 是确定性变量，不是随机变量；自变量之间互不相关。

(2)随机误差项均具有 0 均值和相同的方差，可设 $E(\varepsilon)=0$，$D(\varepsilon)=\sigma^2$，即$\varepsilon\sim N(0, \sigma^2)$。

(3)随机误差项之间不存在序列相关关系；随机误差项与自变量之间不相关。

仿照一元线性回归中的分析步骤，为了估计未知系数 b_0，b_1，b_2，…，b_k，得到经验回归方程：

$$\hat{y} \overset{\text{记为}}{=} b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \tag{3.2}$$

根据多元线性关系式 $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \varepsilon$，对$(y, x_1, x_2, \cdots, x_k)$进行 n 次独立的观测，得容量为 n 的样本值，这里 x_{ik} 表示 x_k 的第 i 次观测值：

$$(y, x_{i1}, x_{i2}, \cdots, x_{ik}) \qquad (i=1, 2, \cdots, n; n>k+1)$$

将上述观测值代入到式(3.1)中得方程组：

$$\begin{cases} y_1 = b_0 + b_1x_{11} + b_2x_{12} + \cdots + b_kx_{1k} + \varepsilon_1 \\ y_2 = b_0 + b_1x_{21} + b_2x_{22} + \cdots + b_kx_{2k} + \varepsilon_2 \\ \cdots \\ y_n = b_0 + b_1x_{n1} + b_2x_{n2} + \cdots + b_kx_{nk} + \varepsilon_n \end{cases} \tag{3.3}$$

如果令：

$$\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)'$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$\boldsymbol{B} = (b_0, b_1, \cdots, b_k)'$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$$

则可得：$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$ (3.4)

这就是式(3.3)的矩阵表示形式，在以后的分析中主要通过它来解决多元分析中的问题。

3.2 参数的最小二乘估计

仿照一元线性回归的参数估计最小二乘方法，对于式(3.1)，系数 b_0，b_1，$\cdots$，b_k 应该取这样的估计值 $\hat{b}_0$，$\hat{b}_1$，$\cdots$，$\hat{b}_k$，能够使得观测值 y_i 与相应的回归值 $b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_kx_{ik}$ 的偏差平方和达到最小，即 b_0，b_1，$\cdots$，b_k 应该取值使函数：

$$Q(b_0, b_1, \cdots, b_k) = \sum_{i=1}^{n} (y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik})^2 \tag{3.5}$$

达到最小。按照多元函数的极值求法，则应使：

$$\begin{cases} \dfrac{\partial Q}{\partial b_0}=-2\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\cdots-b_kx_{ik})=0 \\ \dfrac{\partial Q}{\partial b_1}=-2\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\cdots-b_kx_{ik})x_{i1}=0 \\ \cdots \\ \dfrac{\partial Q}{\partial b_k}=-2\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\cdots-b_kx_{ik})x_{ik}=0 \end{cases} \tag{3.6}$$

上式即为正规方程组。根据上一节的假设，可以将式(3.6)化成如下的形式：

$$\begin{cases} nb_0+\sum\limits_{i=1}^{n}x_{i1}b_1+\cdots+\sum\limits_{i=1}^{n}x_{ik}b_k=\sum\limits_{i=1}^{n}y_i \\ \sum\limits_{i=1}^{n}x_{i1}b_0+\sum\limits_{i=1}^{n}x_{i1}^2b_1+\cdots+\sum\limits_{i=1}^{n}x_{i1}x_{ik}b_k=\sum\limits_{i=1}^{n}x_{i1}y_i \\ \cdots \\ \sum\limits_{i=1}^{n}x_{ik}b_0+\sum\limits_{i=1}^{n}x_{i1}x_{ik}b_1+\cdots+\sum\limits_{i=1}^{n}x_{ik}^2b_k=\sum\limits_{i=1}^{n}x_{ik}y_i \end{cases} \tag{3.7}$$

将式(3.7)表示成矩阵即为：

$$\boldsymbol{X'XB}=\boldsymbol{X'Y}\text{(其中，}\boldsymbol{X'}\text{为}\boldsymbol{X}\text{的转置矩阵)} \tag{3.8}$$

称 $\boldsymbol{X}$ 为结构矩阵，$\boldsymbol{A}=\boldsymbol{X'X}$ 为正规方程组的系数矩阵，$\boldsymbol{X'Y}$ 称为正规方程组的常数项矩阵。如果系数矩阵 $\boldsymbol{A}$ 可逆，那么 $\boldsymbol{B}$ 的估计可以写成如下形式：

$$\hat{\boldsymbol{B}}=(\hat{b}_0,\ \hat{b}_1,\ \cdots,\ \hat{b}_k)'\overset{\text{记为}}{=}(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \tag{3.9}$$

方程 $\hat{y}=\hat{b}_0+\hat{b}_1x_1+\hat{b}_2x_2+\cdots+\hat{b}_kx_k$ 为 k 元经验线性回归方程，简称回归方程。

例 3.1　为了建立国家财政收入回归模型，我们以财政收入 y（单位：10^8 元）为因变量。自变量如下：x_1——工业总产值/10^8 元，x_2——农业总产值/10^8 元，x_3——建筑业总产值/10^8 元，x_4——人口数/10^4 人，x_5——社会商品零售总额/10^8 元，x_6——受灾面积/10^4 hm^2。根据《中国统计年鉴》获得 1978～1990 年统计数据(何晓群，2003)，见表 3.1。

表 3.1　　影响国家财政的各项指标及其取值

年份	y	x_1	x_2	x_3	x_4	x_5	x_6
1978	1 121.1	4 237	1 397	569	96 259	1 558.6	5 076
1979	1 103.3	4 681	1 698	645	97 542	1 800.0	3 937
1980	1 085.2	5 154	1 923	767	98 705	2 140.0	4 453
1981	1 089.5	5 400	2 181	747	100 072	2 350.0	3 979
1982	1 124.0	5 811	2 483	912	101 654	2 570.0	3 313
1983	1 249.0	6 461	2 750	1 035	103 008	2 849.4	3 471
1984	1 501.9	7 617	3 214	1 263	104 357	3 376.4	3 189
1985	1 866.4	9 716	3 619	1 656	105 851	4 350.0	4 437
1986	2 260.3	11 194	4 013	2 038	107 507	4 950.0	4 714
1987	2 386.9	13 813	4 176	2 431	109 300	5 820.0	4 209
1988	2 628.0	18 224	5 865	2 967	111 026	7 440.0	5 087
1989	2 947.9	22 017	6 535	2 834	112 704	8 101.4	4 699
1990	3 244.8	23 851	7 662	3 035	114 333	8 300.1	3 847

由定性分析知，x_1，x_2，x_3，x_4，x_5，x_6 都与变量 y 有较大的相关性。因此，设理论回归模型为：

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_5x_5+b_6x_6+\varepsilon_i \quad (i=1, 2, \cdots, 13)$$

根据表中的统计数据，由最小二乘法计算得到未知参数的估计分别为：

$$\hat{b}_0=460.030\ 1;\ \hat{b}_1=0.078\ 5;\ \hat{b}_2=0.105\ 5;\ \hat{b}_3=0.853\ 2;$$

$$\hat{b}_4=-0.001\ 1;\ \hat{b}_5=-0.007\ 8;\ \hat{b}_6=0.004\ 5$$

则求得 y 关于 x_1，x_2，x_3，x_4，x_5，x_6 的六元线性回归方程为：

$$\hat{y}=460.030\ 1+0.078\ 5x_1+0.105\ 5x_2+0.853\ 2x_3-0.001\ 1x_4-0.007\ 8x_5+0.004\ 5x_6$$

需要注意的是这一回归方程并不理想，回归系数的经济意义不好解释，这里只是作为多元线性回归参数估计的一例，后边我们还要对这一模型作进一步完善。

3.3 回归方程的显著性检验

我们用多元线性回归方程去拟合随机变量 y 与 x_1，x_2，…，x_k 之间的关系，只是根据一些定性分析所作的一些假设。因此，当求出线性回归方程后，还需对回归方程进行显著性检验。实际上，多元线性回归分析相对于一元线性回归分析来说更应该进行显著性检验，因为在一元线性回归分析中，有时我们可以借助由

实验数据建立的散点图来判断拟合的好坏程度，但是当自变量的个数比较多的时候，我们很难建立一个直观的东西来明显地描述自变量和因变量之间的关系。因此，对于多元回归分析，一定要进行显著性检验。

下面简单介绍一下两种常用的统计检验方法，一个是拟合优度检验，另一个是 F 检验。

3.3.1 拟合优度检验

拟合优度检验是检验回归方程对样本观测值的拟合程度。

设 $y=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k+\varepsilon$ 是所求多元线性回归方程，$\hat{y}_i$ 是第 i 个样本点(x_{i1}，x_{i2}，…，x_{ik})上的回归值。我们用 $y_i-\bar{y}$ 表示 y 的第 i 个观测值与 y 的样本平均值的偏差。因为观测值 y_1，y_2，…，y_n 之间的差异是由自变量取值的不同和其他随机因素两个方面引起的，为了知道这两者之间哪一个是主要的，我们有必要把 y 的总偏差分解，于是总偏差平方和可分解为 $S_{残}$ 和 $S_{回}$ 两部分。即：

$$\begin{aligned}S_{总}&=\sum_{i=1}^{n}(y_i-\bar{y})^2\\&=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\\&=S_{残}+S_{回}\end{aligned}$$

其中，$\hat{y}_i$ 称为回归值，它是由回归方程计算出的因变量在第 i 个样本点上的取值。$S_{残}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 称为残差平方和，$S_{回}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 称为回归平方和，交叉项 $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$ 为零，具体原因同一元线性回归中的分析，见第2章2.2节。

残差平方和反映了自变量 x 对因变量 y 的线性影响之外的一切因素(包括 x 对 y 的非线性影响和测量误差等)对因变量 y 的作用。回归平方和反映了总偏差平方和中由于 x 与 y 的线性关系而引起因变量 y 变化的部分。

由 $S_{残}$ 和 $S_{回}$ 的意义可知，一个好的回归方程，它应该较好地拟合样本的观测值。总的偏差平方和中，回归平方和所占的比例越大，则线性回归效果越显著；残差平方和所占比例越大，则线性回归效果就越不显著。于是定义如下的系数来反映自变量与因变量之间的线性回归效果显著程度：

$$r^2=\frac{S_{回}}{S_{总}};\quad r=\sqrt{\frac{S_{回}}{S_{总}}} \tag{3.10}$$

其中 r^2 称为样本决定系数，r 称为 y 关于 x_1，x_2，…，x_k 样本的复相关系数。与一元线性回归方程中曾定义过的相关系数 r 一样，在多元线性回归的实际应用中，人们通常用复相关系数 r 来表示回归方程对原有数据拟合程度的好坏，衡量作为一个整体的 x_1，x_2，…，x_k 与 y 线性关系的显著程度。

如果回归方程完全拟合样本观测值，则：$y_i-\hat{y}_i=0\quad(i=1,2,\cdots,n)$。

由此：
$$S_{残}=\sum_{i=1}^{n}(y_i-\hat{y})^2=0;\ r^2=\frac{S_{回}}{S_{总}}=1-\frac{S_{残}}{S_{总}}=1$$

完全拟合是一种极端情况，这在实际问题的研究中不大可能出现，即 r^2 不可能等于 1。很容易理解，如果 r^2 越接近于 1，回归方程的拟合优度越高。

类似于一元线性回归分析，当给定显著性水平 α 时，由式(3.10)计算出 r 值，再根据相关系数检验表（附表 2），查出 $r_\alpha(n-2)$的值。

如果 $r>r_\alpha(n-2)$，则可以认为多元线性回归是显著的；如果 $r\leqslant r_\alpha(n-2)$，则可以认为多元线性回归不显著，自变量和因变量之间的关系不能用线性关系来描述。

根据例 3.1，给定显著性水平 $\alpha=0.05$，则：

$$S_{回}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=7\ 320\ 924.00$$

$$S_{总}=\sum_{i=1}^{n}(y_i-\bar{y})^2=7\ 439\ 339.00$$

$$S_{残}=S_{总}-S_{回}=118\ 415.60$$

则样本决定系数和复相关系数分别为：

$$r^2=\frac{S_{回}}{S_{总}}=\frac{7\ 320\ 924.00}{7\ 439\ 339.00}=0.984\ 1;\ r=0.992\ 0$$

在显著性水平 $\alpha=0.05$ 下，$r_\alpha(n-2)=r_\alpha(11)=0.553$。

由于 $r=0.992\ 0>0.553=r_\alpha(n-2)$，所以认为财政收入的回归方程拟合优度很高。

在实际应用中，决定系数 r^2 到底多大才算通过了拟合优度检验，要根据具体情况来定。在此需要指出的是，拟合优度并不是检验模型优劣的唯一标准，有时为了使模型在结构上有较合理的经济解释，$r^2=0.7$ 左右我们也对模型给以肯定。实际上，在多元回归分析中，r^2 与回归方程中自变量的数目以及样本容量 n 有关。当样本容量的个数与自变量的个数接近时，r^2 易接近于 1，其中隐含着一些虚假成分。因此，我们在使用 r^2 决定模型优劣时还需要慎重。

3.3.2 *F* 检验

拟合优度检验主要是利用观测值和实际值之间的误差大小来判断拟合的显著

程度。下面介绍的 F 检验，主要是通过对回归系数的显著性检验来判断多元线性回归分析的拟合程度。实际上，对回归方程显著性检验，就是要看自变量 x_1，x_2，…，x_k 从整体上对随机变量 y 是否有明显的影响。为此，可提出假设：

$$H_0: b_1=b_2=\cdots=b_k=0$$

如果 H_0 被接受，则表明随机变量 y 与 x_1，x_2，…，x_k 之间的关系由线性回归模型表示不合适。类似一元线性回归检验，可建立起多元线性回归方程显著性检验的 F 统计量。

$$F=\frac{S_{回}/k}{S_{残}/(n-k-1)}\sim F(k, n-k-1) \tag{3.11}$$

于是，可利用 F 统计量对回归方程的总体显著性进行检验。对于给定的数据 $(y, x_{i1}, x_{i2}, \cdots, x_{ik})(i=1, 2, \cdots, n; n>k+1)$，计算出 $S_{回}$ 和 $S_{残}$，进而得到 F 的值，再由给定的显著性水平 α，查 F 分布临界值表（附表5），得临界值 $F_\alpha(k, n-k-1)$。

当 $F\geqslant F_\alpha(k, n-k-1)$，则拒绝假设 H_0，认为在显著性水平 α 下，y 对 x_1，x_2，…，x_k 有显著的线性关系，也即回归方程是显著的；反之，则认为回归方程不显著。例如，给定 $\alpha=0.05$，如果 $F\geqslant F_\alpha(k, n-k-1)$，则在5%的显著性水平上，应该拒绝 H_0，也即 y 对 x_1，x_2，…，x_k 有显著的线性关系。

根据例3.1，在给定显著性水平 $\alpha=0.05$ 的情况下，可以计算：

$$F=\frac{S_{回}/k}{S_{残}/(n-k-1)}=\frac{7\ 320\ 924.00/6}{118\ 415.60/6}=61.824\ 0$$

当 $\alpha=0.05$ 时，根据 F 分布临界值表（附表5），查得 $F_\alpha(k, n-k-1)=F_\alpha(6, 6)=4.28$，由于 $F=61.824\ 0>4.28$，所以可以认为财政收入的回归方程拟合优度很高，这与拟合优度检验的结果一致。

3.4 回归系数的显著性检验

在多元线性回归中，回归方程显著并不意味着每个自变量对 y 的影响都显著。因此应从回归方程中剔除那些次要的、可有可无的变量，重新建立更为简单的回归方程，所以就需要对每个自变量进行显著性检验。

显然，如果某个自变量 x_i 对 y 的作用不显著，那么在回归模型中，它的系数 b_i 就可以取值为零。因此，检验变量 x_i 是否显著等价于检验假设：$H_{0i}: b_i=0(i=1, 2, \cdots, k)$。

如果没有足够的理由否定假设 H_0，则通常认为 x_i 不显著；如果拒绝假设，则 x_i 显著。

容易证明，当 H_{0i}成立时：$\frac{b_i}{\sqrt{c_{ii}\sigma^2}} \sim N(0,\ 1)$。

在上面假设下，可采用 $F=\frac{\hat{b}_i^2/c_{ii}}{S_{残}/(n-k-1)} \sim F(1,\ n-k-1)$进行检验。其中，$c_{ii}$是矩阵$(\boldsymbol{X'X})^{-1}$对角线上第 i 个元素，可用 F 来检验b_i 是否为零，即x_i 对 y 的影响是否显著。

根据例 3.1，我们已经看到回归方程：

$$\hat{y}=462.030\ 1+0.078\ 5x_1+0.105\ 5x_2+0.853\ 2x_3-0.001\ 1x_4-0.307\ 8x_5+0.044\ 5x_6$$

是十分显著的，然而这种显著是 x_1，x_2，x_3，x_4，x_5，x_6 作为一个整体变量对因变量 y 产生的十分显著的影响。每一个自变量 $x_i(i=1,\ 2,\ \cdots,\ 6)$是否对 y 有显著影响呢？这就需对假设 H_0：$b_i=0(i=1,\ 2,\ \cdots,\ 6)$进行检验。

利用 Matlab 软件计算，得关于 $b_i(i=1,\ 2,\ \cdots,\ 6)$的 F 统计量 $F_i(i=1,\ 2,\ \cdots,\ 6)$，如下：

$$F_1=1.076\ 9,\ F_2=0.297\ 4,\ F_3=2.911\ 9,$$
$$F_4=0.035\ 2,\ F_5=0.659\ 7,\ F_6=0.160\ 8$$

查 F 分布临界值表（附表 5），得：

$$F_\alpha(1,\ n-k-1)=F_\alpha(1,\ 6)=5.99$$

上述 $F_i(i=1,\ 2,\ \cdots,\ 6)<F_\alpha(1,\ 6)=5.99$，即说明每一个 x_i 单独对因变量 y 无显著性影响。这个例子说明，尽管回归方程通过了显著性检验，但也会出现某些单个变量 x_i 对 y 并不显著的情况，这也说明变量之间有一定的交互作用，后面将会进一步看到不同变量组合在一起建立方程的效果是不一样的。为了使模型简化些，我们可以将对因变量 y 影响不显著的变量剔除，然后重新利用最小二乘法建立回归方程。当有多个自变量对因变量 y 无显著性影响时，由于 b_i 的各分量间的相关性，不能一次取消掉所有不显著的变量。原则上每次只剔除一个变量，先剔除其中 F 值最小的一个变量，然后再对求得的新回归方程进行检验，如果不显著，再剔除变量，直到保留的变量都对 y 有显著性影响为止。也可根据对问题的定性分析选择 F 值较小的变量先剔除。

例 3.2 财政收入一项中，x_4 为人口数，$F_4=0.035\ 2<F_\alpha(1,\ 6)=5.99$，对财政收入的影响相对较小，故我们剔除 x_4，用最小二乘法建立新的回归方程：

$$\hat{y}=331.95+0.085\ 6x_1+0.106\ 3x_2+0.878\ 4x_3-0.342\ 8x_5+0.056\ 80x_6$$

此时，样本决定系数 $r^2=0.992\ 0$，复相关系数 $r=0.984\ 0$，$F=86.040\ 8$，经查 F 分布临界值表（附表 5）知，$F_\alpha(1,\ n-k-1)=F_\alpha(1,\ 7)=5.59$，显然这个线性回归方程是高度显著的。

值得说明的是，多元线性回归方程中并非自变量越多越好，原因是由于自变量越多剩余标准差可能变大，同时也增加了收集资料的难度。故需寻求“最佳”回归方程，逐步回归分析是寻求“较佳”回归方程的一种方法。有关逐步回归分析方法详见有关参考文献。

3.5 Matlab 语言在多元回归中的应用

回归分析是数理统计中最常用的方法之一，一般用最小二乘法确定回归方程中的系数，其矩阵计算过程颇为复杂，而用 Matlab 实现则使问题大大简化。Matlab 中有四个函数可以用于回归分析和拟合：polyfit（x，y，n），leastsp（/function/，x），regress(y，x)和 pinv(A) * y。polyfit（x，y，n）只能用于多项式线性回归，leastsp(/function/，x)可用来作非线性回归，regress(y，x)可用于多元线性回归，pinv(A) * y 可用于求解线性方程组。

例 3.3　在一定的温度下，饱和醇类化合物的拓扑指数及保留指数值见表 3.2。求饱和醇类化合物拓扑指数与保留指数之间的关系。

表 3.2　**醇类化合物拓扑指数及保留指数数据表**

编号	醇	保留指数(y)			拓扑指数(A)			
		SE-30	OV-3	OV-7	$^{1}\chi$	$^{2}\chi_{P}$	$^{2}\chi-^{2}\chi^{V}$	C_{OH}
1	1-丁醇	649	673	701	2.424	0.703	0.275	0.709
2	1-己醇	857	882	909	3.415	1.208	0.277	0.709
3	1-庚醇	961	986	1 010	3.915	1.456	0.277	0.709
4	2-丁醇	587	608	634	2.271	0.817	0.545	0.579
5	2-戊醇	687	710	734	2.769	0.865	0.543	0.575
6	3-戊醇	687	709	734	2.807	1.393	0.450	0.575
7	3-己醇	784	806	829	3.307	1.477	0.450	0.575
8	3-庚醇	885	907	927	3.807	1.745	0.450	0.575
9	4-庚醇	881	905	925	3.809	1.564	0.452	0.579
10	2-甲基-2-丁醇	629	653	675	2.562	1.061	0.749	0.501

根据表 3.2 中的数据建立相应的 M 文件，M 文件中输入以下代码（说明：以下代码格式为文件中真实格式）。

```
A= [1  2.424  0.703  0.275   0.709
    1  3.415  1.208  0.277   0.709
    1  3.915  1.456  0.277   0.709
    1  2.271  0.817  0.545   0.579
    1  2.769  0.865  0.543   0.575
    1  2.807  1.393  0.45    0.575
    1  3.307  1.477  0.45    0.575
    1  3.807  1.745  0.45    0.575
    1  3.809  1.564  0.452   0.579
    1  2.562  1.061  0.749   0.501];
y=  [649 673 701
     857 882 909
     961 986 1010
     587 608 634
     687 710 734
     687 709 734
     784 806 829
     885 907 927
     881 905 925
     629 653 675];
```

$\hat{b}$=pinv(A) * y

得到运行的结果为：

$\hat{b}$=

```
  -322.4033   -334.0991   -319.6683
   195.3720    196.9256    192.2776
    17.8680     17.4884     19.9566
   159.0413    175.3433    175.4272
   628.9059    667.1071    701.3615
```

所以，求得的拟合方程为：

$y_{SE\text{-}30}=-322.4033+195.3720\,{}^{1}\chi+17.8680\,{}^{2}\chi_{P}+159.0413({}^{2}\chi-{}^{2}\chi^{V})+628.9059C_{OH}$

$y_{OV\text{-}3}=-334.0991+196.9256\,{}^{1}\chi+17.4884\,{}^{2}\chi_{P}+175.3433({}^{2}\chi-{}^{2}\chi^{V})+667.1071C_{OH}$

$$y_{OV\text{-}7}=-319.6683+192.2776\,{}^{1}\chi+19.9566\,{}^{2}\chi_{P}+175.4272({}^{2}\chi-{}^{2}\chi^{V})+701.3615C_{OH}$$

由此可知，Matlab 作为新一代科学和工程计算语言，其简洁、易操作性是其他类似软件所不能比拟的，它应该是我们进行环境科学与环境工程计算的首选工具。鉴于 Matlab 强大的计算能力和优越性，在下一节的案例分析中，我们都将采取基于 Matlab 编程的方法来求解有关问题。

3.6 环境应用

例 3.4 对某河的主要排污沟进行调查监测，详见下表。在监测中发现该水域的总悬浮物(suspended sediment，简称 SS)、COD、BOD 成正相关，试建立 BOD 与 SS、COD 的关系，对此进行多元线性回归分析。

表 3.3 监测结果

编号	SS	COD	BOD
1	413	45.60	13.59
2	363	37.72	12.78
3	803	70.65	26.29
4	730	81.47	23.97
5	823	90.83	28.09
6	589	58.95	18.06
7	523	50.39	17.84
8	674	61.50	22.18
9	984	107.17	34.07
10	1 369	130.79	45.89

根据表 3.3 中的数据建立相应的 M 文件，M 文件中输入以下代码（说明：以下代码格式为文件中真实格式）。

```
A=[1  413  45.60;  1  363   37.72
   1  803  70.65;  1  730   81.47
   1  823  90.83;  1  589   58.95
   1  523  50.39;  1  674   61.50
   1  984  107.17; 1  1369  130.79];
y=[13.59;12.78;26.29;23.97;28.09;18.06;17.84;22.18;34.07;45.89];
```

```
C=A' A=
  1.0e+006 *
     [0.0000     0.0073     0.0007
      0.0073     6.0745     0.6105
      0.0007     0.6105     0.0618]
```

$\hat{b}$=pinv(A) * y

$\hat{y}$=A * $\hat{b}$

得到运行结果为：

$\hat{b}$=[−0.5584；　0.0293；　0.0483]

$\hat{y}$=[13.7335；　11.8893；　26.3602；　24.7452；　27.9197
　19.5304；　17.1850；　22.1419；　33.4217；　45.8330]

由此得到回归方程为：

$$BOD=-0.5584+0.0293SS+0.0483COD$$

回归方程的 F 检验列于下表：

$$F_{0.01}(v_1, v_2)=F_{0.01}(2, 7)=9.55$$

$$r^2=0.9951$$

表 3.4　　回归方程的检验结果

误差来源	平方和	自由度 v	r^2	F	显著性
回归	908.121 7	2	0.995 1	711.895 9	高度显著
残差	4.464 7	7			
总和	912.586 4				

可见该回归方程的显著性很好，说明该水域的 BOD 指标受到 SS 和 COD 的影响很大。

例 3.5　洛河污染分析

考察洛河在安乐窝—十方院渡口段河水受污染情况。考察指标 y 表示 BOD 浓度。而 BOD 浓度 y 可能与以下几个因素有关(卢崇飞等，1988)：x_1：初始断面 BOD 浓度；x_2：初始断面氧亏浓度 C_0；x_3：水温 T；x_4：河流流量 q；x_5：排污口流量 Q；x_6：污水 BOD 浓度 l；x_7：流过该河段所需时间 t。

表 3.5　　监测结果

编号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	6.88	−0.25	27.0	674 784	11 232	477	0.083	9.35
2	6.08	−2.21	27.5	477 792	11 232	193	0.083	12.30
3	2.14	−3.04	26.0	477 792	11 232	404	0.083	15.60
4	5.02	−0.75	26.0	856 224	11 232	363	0.073	5.88
5	7.89	−2.26	26.0	856 224	11 232	363	0.069	6.34
6	2.38	−1.65	15.0	1 490 400	15 552	428	0.104	4.00
7	1.86	−1.35	15.8	1 490 400	15 552	428	0.104	3.76
8	1.02	−2.12	17.1	1 494 720	13 824	428	0.104	3.98
9	1.22	−1.92	17.5	1 494 720	13 824	428	0.104	3.98
10	0.90	−0.27	17.0	3 628 800	9 936	202	0.104	2.78
11	1.58	−0.09	17.0	3 628 800	9 936	202	0.104	1.88
12	2.78	−1.17	13.5	3 265 920	9 936	114	0.104	2.56
13	2.10	−1.30	13.5	3 265 920	9 936	114	0.104	2.72
14	2.32	−0.60	14.5	3 646 080	8 640	57.3	0.104	1.64
15	1.96	−0.60	14.5	3 646 080	8 640	57.3	0.104	2.36

根据表 3.5 中的数据建立相应的 M 文件，M 文件中输入以下代码（说明：以下代码格式为文件中真实格式）。

```
A= [1  6.88  -0.25  27.0  674784   11232  477   0.083
    1  6.08  -2.21  27.5  477792   11232  193   0.083
    1  2.14  -3.04  26.0  477792   11232  404   0.083
    1  5.02  -0.75  26.0  856224   11232  363   0.073
    1  7.89  -2.26  26.0  856224   11232  363   0.069
    1  2.38  -1.65  15.0  1490400  15552  428   0.104
    1  1.86  -1.35  15.8  1490400  15552  428   0.104
    1  1.02  -2.12  17.1  1494720  13824  428   0.104
    1  1.22  -1.92  17.5  1494720  13824  428   0.104
    1  0.90  -0.27  17.0  3628800   9936  202   0.104
    1  1.58  -0.09  17.0  3628800   9936  202   0.104
    1  2.78  -1.17  13.5  3265920   9936  114   0.104
    1  2.10  -1.30  13.5  3265920   9936  114   0.104
    1  2.32  -0.60  14.5  3646080   8640  57.3  0.104
    1  1.96  -0.60  14.5  3646080   8640  57.3  0.104];
```

```
y= [9.35
    12.30
    15.60
    5.88
    6.34
    4.00
    3.76
    3.98
    3.98
    2.78
    1.88
    2.56
    2.72
    1.64
    2.36];
b̂=pinv(A) * y
  ŷ=A * b̂
```

得到运行结果为：

```
b̂=-9.2100
   -0.3179
   -1.1490
    0.6117
   -0.0000
   -0.0009
   -0.0027
  202.4915
```

```
ŷ=8.5835
  12.7210
  13.4341
  6.9164
  6.9289
  2.3714
  2.6814
  6.2229
  6.1743
  2.3401
  1.9171
  1.8898
  2.2553
  2.2898
  2.4042
```

由此得到回归方程为：

$$\hat{y}=-9.2100-0.3179x_1-1.1490x_2+0.6117x_3-0.0000x_4-0.0009x_5-0.0027x_6+202.4915x_7$$

回归方程的检验列于下表：

表 3.6　回归方程的检验结果

误差来源	平方和	自由度	r^2	显著性
回归	212.875 4	7	0.907 0	显著
残差	21.822 4	7		
总和	234.697 8			

【思考题 3】

1. 多元线性回归模型有哪些基本假定？

2. 表 3.7 是某湖区历年实测的湖水污染物 COD 浓度与相应的环境自然经济资料。根据专业经验分析，认为湖泊水质污染浓度的高低，一方面取决于沿湖地区工农业生产发展所排放的污染物质的数量，另一方面与湖泊水文状况有关。试用多元线性回归分析方法研究湖泊水质污染预测问题，说明其规律，并分析各因子的贡献大小(陈玉成等，1998)。

表3.7　湖泊影响因子及其监测值

项目 年份	COD浓度 $y/(\mathrm{mg}\cdot\mathrm{L}^{-1})$	农业产量 $x_1/(5\times10^7\ \mathrm{kg})$	工业总产值 $x_2/10^8$ 元	湖泊水位 x_3/m
1960	2.50	0.25	4.00	3.17
1975	2.63	0.92	21.10	3.24
1976	3.15	0.87	29.10	3.02
1977	2.52	0.60	33.00	3.24
1978	4.06	0.63	37.50	2.63
1979	3.72	0.65	42.40	2.80
1980	2.82	0.42	49.25	3.85
1981	3.31	0.40	50.00	2.97

3. 根据统计资料显示(表3.8)，影响铁路旅客周转量的可能因素有：铁路运营里程、铁路客车数量、公路通车里程(有等级公路)、公路客车数量。试建立铁路旅客周转量的多元线性回归模型。

表3.8　影响铁路旅客周转量的可能因素及其测定值

年份	铁路旅客周转量 $/(10^8$ 人·km)	铁路客车数量 /辆	铁路运营里程 $/10^4$ km	公路通车里程 $/10^4$ km	公路客车 $/10^4$ 辆
1986	2 583	22 138	5.25	63.77	96.61
1987	1 840	23 474	5.26	66.84	111.46
1988	3 257	24 917	5.28	69.73	130.38
1989	3 034	26 304	5.32	71.69	146.43
1990	2 610	27 261	5.34	74.11	162.19
1991	2 825	27 612	5.34	76.47	185.24
1992	3 148	28 464	5.36	78.69	226.16
1993	3 479	29 645	5.38	82.21	285.98
1994	3 633	31 268	5.40	86.14	349.74
1995	3 543	32 663	5.46	91.08	417.90
1996	3 322	33 778	5.67	94.81	488.02
1997	3 544	34 346	5.76	99.75	580.56

4. 表 3.9 给出了出厂水浊度(y)与净化药剂投加量（采用耗矾率 x_1）、原水浊度（x_2）之间的关系。

(1)计算出 y，x_1，x_2 的相关系数矩阵；

(2)求 y 与 x_1，x_2 的二元线性回归方程；

(3)对所求得的回归方程作拟合优度检验，并对回归方程和每一个回归系数作显著性检验。

表 3.9　　统计数据表

时　间	耗矾率/(kg・$(10^3\text{t})^{-1}$)	原水浊度	出厂水浊度
6 月 2 日	14.2	57	0.17
6 月 3 日	13.6	52	0.23
6 月 4 日	13.8	47	0.24
6 月 5 日	15.2	47	0.18
6 月 6 日	15.0	45	0.18
6 月 7 日	14.5	41	0.21
6 月 8 日	15.1	42	0.21
6 月 9 日	14.8	39	0.20
6 月 10 日	15.2	44	0.18
6 月 11 日	14.6	58	0.13
6 月 12 日	15.0	72	0.13
6 月 13 日	15.5	88	0.14
6 月 14 日	15.3	87	0.14
6 月 15 日	15.3	86	0.13
6 月 16 日	14.8	112	0.14

5. 试用多元线性回归模型预测一个实际的环境问题。

【参考文献】

[1] 何晓群. 现代统计分析方法与应用 [M]. 北京：中国人民大学出版社，2003.

[2] 卢崇飞，高惠璇，叶文虎. 环境数理统计学应用及程序 [M]. 北京：高等教育出版社，1988.

[3] 陈玉成，吕宗清，李章平. 环境数学分析 [M]. 重庆：西南师范大学出版社，1998.

第 4 章　环境系统聚类分析

"物以类聚"，分类是许多学科领域的重要内容。日常生活和实践中，我们常常把所接触、研究的对象，按照它们的性质、用途等分成几类。例如，地质勘探中，要按照矿石标本的颜色、比重和化学成分等特性将矿石分成很多不同的类别；气象学中，常需要按照大气环流的不同，将大气形式分成若干环流型；在水环境评价中，常根据水质污染水平的不同，把水域分成若干类型。随着环境科学自动化分析技术的迅速普及，环境问题如何归类和分析已成为环境科学的一项重要课题。在环境数据分类中，这种按确定的标准对客观事物进行分级、分类的数学方法称为环境聚类分析。

系统聚类分析(hierarchical cluster analysis)是环境聚类分析中应用较广泛的一种方法，凡是具有数值特征的变量和样本都可以采用系统聚类分析法。其基本原理是根据样本自身的属性，用数学方法按照某种相似性或差异性指标，定量地确定样本之间的亲疏关系，并按这种亲疏关系程度对样本进行聚类。

本章的主要内容是：

- 聚类分析概述；
- 聚类要素的数据处理；
- 距离和相似系数的计算；
- 系统聚类分析的常用方法；
- 环境应用。

4.1　聚类分析概述

聚类(cluster)就是按照事物间的相似性进行区分和分类的过程，在这一过程中没有教师指导，因此是一种无监督的分类。聚类分析(cluster analysis)又称点群分析、群分析、簇分析等，它是研究样本(或指标)分类问题的一种统计分析方法。聚类分析起源于分类学，在古老的分类学中，人们主要依靠经验和专业知识来实现分类，很少利用数学工具进行定量的分类。随着人类科学技术的发展，对分类的要求越来越高，以致有时仅凭经验和专业知识难以确切地进行分类，于是人们逐渐地把数学工具引用到了分类学中，形成了数值分类学，之后又将多元分

析的技术引入到数值分类学中，形成了聚类分析。聚类分析内容非常丰富，有系统聚类法、有序样本聚类法、动态聚类法、图论聚类法、聚类预报法等。

聚类分析的基本思想是认为我们所研究的样本或指标(变量)之间存在着某种程度的相似性(亲疏关系)。首先，将要归类的 n 个样本各自看成一类，然后按事先规定好的方法计算各类之间的归类指数(如某种相关系数或距离)，根据指数大小衡量两两之间的密切程度，将关系最密切的两类并成一类，其余不变，即得 $n-1$ 类；又按事先规定的方法重新计算各类之间的归类指数(仍为某种相关系数或距离)，又将关系最密切的两类并成一类，其余不变，即得 $n-2$ 类；如此进行下去，每归类一次都减少一类，直至最后，n 个变量都归成一类为止。这一归类过程可以用一张聚类图形象地表示出来。聚类分析诸多方法中最常用、最基本的一种是系统聚类分析。

通常还将聚类分析根据分类对象的不同分为 Q 型和 R 型两大类。Q 型是对样本进行分类处理(如解剖学上依据骨骼的形状和大小等，不仅可以区别样本是人还是猿，还可以区别性别、年龄等)，R 型是对变量进行分类处理(如在儿童的生长发育研究中，把以形态学为主的指标归于一类，以机能为主的指标归于另一类等)。常用的聚类统计量有距离系数和相似系数两类。距离系数一般用于对样本分类，而相似系数一般用于对变量聚类。

Q 型聚类分析的特点是：(1)可以综合利用多个变量的信息对样本进行分类；(2)分类结果是直观的，聚类谱系图非常清晰地表达出其数值分类的结果。

R 型聚类分析的特点是：(1)不但可以了解个别变量之间的亲疏程度，而且可以了解各个变量组合之间的亲疏程度；(2)根据变量的分类结果以及它们之间的关系，可以选择主要变量进行回归分析或 Q 型聚类分析。

4.2 聚类要素的数据处理

在系统聚类分析中，聚类要素的选择是十分重要的，它直接影响分类结果的准确性和可靠性。在环境科学研究中，被聚类的对象通常是多个要素构成的，不同要素往往具有不同的单位和量纲，因而其数值的差异可能很大，这就会对分类结果产生影响。因此，当分类对象确定后，在进行系统聚类分析之前，还要对聚类要素进行数据处理。值得注意的是，聚类要素数据矩阵中，一般行表示样本，列表示变量（指标）。

在聚类分析中，常用的聚类要素的数据处理方法主要有：总和标准化、标准差标准化、极大值标准化、极差标准化等。

例 4.1　以长江流域水环境数据为例，1993 年 1 月份 6 个站点水环境监测指标实测值如表 4.1 所示。

表 4.1　**1993 年 1 月份各站点水环境监测指标实测值　单位：mg/L**

各站点	指标				
	溶解氧	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
攀枝花	10.0	0.8	2.0	0.10	0.003
高　场	10.5	1.3	1.8	0.16	0.002
津　市	10.4	1.9	1.2	0.16	0.003
长　沙	8.8	2.3	1.1	0.72	0.002
中山桥	13.0	3.5	2.9	0.30	0.019
宣　城	13.4	2.3	2.4	0.02	0.005

(1)总和标准化：分别求出各聚类要素所对应的数据的总和，以各要素的数据除以该要素数据的总和，即：

$$x_{ij}' = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \quad (i=1,\ 2,\ \cdots,\ m;\ j=1,\ 2,\ \cdots,\ n) \tag{4.1}$$

且
$$\sum_{i=1}^{m} x_{ij}' = 1 \quad (j=1,\ 2,\ \cdots,\ n)$$

其中，x_{ij}'为总和标准化后的数据；x_{ij} 表示第 i 样本的第 j 个指标。

表 4.1 数据经总和标准化处理后，得到如表 4.2 中所列的数据。

在 Matlab 环境下程序（说明：文字格式为程序中真实格式）为：

```
x=[10.0  0.8  2.0  0.10  0.003; 10.5  1.3  1.8  0.16  0.002;
   10.4  1.9  1.2  0.16  0.003; 8.8  2.3  1.1  0.72  0.002;
   13.0  3.5  2.9  0.30  0.019; 13.4  2.3  2.4  0.02  0.005]
xx1=zeros(6,5);
for i=1:6
    for j=1:5
        t=sum(x,1);
xx1 (i,j)=x(i,j)/t(j);
end
end
xx1
```

其中，x 为标准化前的数据，$xx1$ 为总和标准化处理后的数据。

表 4.2　　　　总和标准化变换结果

各站点	指标				
	溶解氧	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
攀枝花	0.151 3	0.066 1	0.175 4	0.068 5	0.088 2
高　场	0.158 9	0.107 4	0.157 9	0.109 6	0.058 8
津　市	0.157 3	0.157 0	0.105 3	0.109 6	0.088 2
长　沙	0.133 1	0.190 1	0.096 5	0.493 2	0.058 8
中山桥	0.196 7	0.289 3	0.254 4	0.205 5	0.558 8
宣　城	0.202 7	0.190 1	0.210 5	0.013 7	0.147 1

（2）标准差标准化，即：

$$x_{ij}'=\frac{x_{ij}-\bar{x}_j}{s_j}\quad(i=1,2,\cdots,m;\ j=1,2,\cdots,n)\tag{4.2}$$

其中，

$$\bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij},\quad s_j=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2}$$

且

$$\bar{x}_j'=\frac{1}{m}\sum_{i=1}^{m}x_{ij}'=0,\quad s_j'=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ij}'-\bar{x}_j')^2}=1$$

其中，x_{ij}'，x_{ij} 含义同式(4.1)；$\bar{x}_j$ 为第 j 个指标的平均值；s_j 为第 j 个指标的标准差。通过变换处理后，每列数据的平均值为 0，方差为 1。使用标准差标准化处理后，在抽样样本改变时，它仍保持相对稳定性。

表 4.1 中的数据经标准差标准化处理后，得如表 4.3 中所列的数据。在 Matlab 环境下程序（说明：文字格式为程序中真实格式）为：

```
x=[10.0  0.8  2.0  0.10  0.003; 10.5  1.3  1.8  0.16  0.002;
   10.4  1.9  1.2  0.16  0.003; 8.8  2.3  1.1  0.72  0.002;
   13.0  3.5  1.9  0.30  0.019; 13.4  2.3  2.4  0.02  0.005]
y=std(x,1);
z=mean(x,1);
xx2=zeros(6,5);
for j=1:5
    for i=1:6
        xx2 (i,j)=(x(i,j)-z(j))/y(j);
```

```
end
end
xx2
```

其中，x 为标准化前的数据，$xx2$ 为标准差标准化处理后的数据。

表 4.3 标准差标准变换结果

各站点	指标				
	溶解氧	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
攀枝花	−0.618 6	−1.425 9	0.158 1	−0.625 8	−0.441 1
高　场	−0.314 4	−0.839 9	−0.158 1	−0.363 9	−0.606 4
津　市	−0.375 2	−0.136 7	−1.106 8	−0.363 9	−0.441 1
长　沙	−1.348 7	0.332 1	−1.264 9	2.081 2	−0.606 4
中山桥	1.206 7	1.738 4	1.581 1	0.247 4	2.205 3
宣　城	1.450 1	0.332 1	0.790 6	−0.975 1	−0.110 3

(3)极大值标准化，即：

$$x_{ij}'=\frac{x_{ij}}{\max\limits_{i}\{x_{ij}\}}\quad(i=1,2,\cdots,m;\ j=1,2,\cdots,n)\tag{4.3}$$

其中，x_{ij}'，x_{ij} 含义同式(4.2)；$\max\limits_{i}\{x_{ij}\}$ 为 i 样品的第 j 个指标中的最大值。经过这种标准化所得的新数据，各要素的极大值为1，其余各数值小于1。

表4.1数据经极大值标准化处理后，得如表4.4中所列的数据。

在 Matlab 环境下程序（说明：文字格式为程序中真实格式）为：

```
x=[10.0  0.8  2.0  0.10  0.003;   10.5  1.3  1.8  0.16  0.002;
   10.4  1.9  1.2  0.16  0.003;    8.8  2.3  1.1  0.72  0.002;
   13.0  3.5  1.9  0.30  0.019;   13.4  2.3  2.4  0.02  0.005]
a=max(x,[  ],1);
xx3=zeros(6,5);
for j=1:5
    for i=1:6
        xx3 (i,j)=x(i,j)/a(j);
end
end
xx3
```

其中，x 为标准化前的数据，$xx3$ 为极大值标准化处理后的数据。

表 4.4　　极大值标准化变换结果

各站点	指标				
	溶解氧	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
攀枝花	0.746 3	0.228 6	0.689 7	0.138 9	0.157 9
高　场	0.783 6	0.371 4	0.620 7	0.222 2	0.105 3
津　市	0.776 1	0.542 9	0.413 8	0.222 2	0.157 9
长　沙	0.656 7	0.657 1	0.379 3	1.000 0	0.105 3
中山桥	0.970 1	1.000 0	1.000 0	0.416 7	1.000 0
宣　城	1.000 0	0.657 1	0.827 6	0.027 8	0.263 2

(4)极差的标准化，即：

$$x'_{ij}=\frac{x_{ij}-\min\limits_{i}\{x_{ij}\}}{\max\limits_{i}\{x_{ij}\}-\min\limits_{i}\{x_{ij}\}}\quad (i=1,\ 2,\ \cdots,\ m;\ j=1,\ 2,\ \cdots,\ n)\ (4.4)$$

经过这种标准化所得的新数据，各要素的极大值为 1，极小值为 0，其余的数值均在 0 与 1 之间。

表 4.1 数据经标准差标准化处理后，得如表 4.5 中所列的数据。

在 Matlab 环境下程序（说明：文字格式为程序中真实格式）为：

```
x=[10.0  0.8  2.0  0.10  0.003; 10.5  1.3  1.8  0.16  0.002;
   10.4  1.9  1.2  0.16  0.003; 8.8  2.3  1.1  0.72  0.002;
   13.0  3.5  1.9  0.30  0.019; 13.4  2.3  2.4  0.02  0.005]
a=max (x,[  ],1);
b=min(x,[  ],1);
xx4=zeros(6,5);
for j=1:5
   for i=1:6
      xx4 (i,j)=(x(i,j)-b(j))/(a(j)-b(j));
end
end
xx4
```

其中，x 为标准化前的数据，$xx4$ 为极差标准化处理后的数据。

表4.5 极差标准化变换结果

各站点	指标				
	溶解氧	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
攀枝花	0.260 9	0.000 0	0.500 0	0.114 3	0.058 8
高　场	0.369 6	0.185 2	0.388 9	0.200 0	0.000 0
津　市	0.347 8	0.407 4	0.055 6	0.200 0	0.058 8
长　沙	0.000 0	0.555 6	0.000 0	1.000 0	0.000 0
中山桥	0.913 0	1.000 0	1.000 0	0.400 0	1.000 0
宣　城	1.000 0	0.555 6	0.722 2	0.000 0	0.176 5

此外，还有中心化标准化、对数标准化、平方根标准化、立方根标准化等。立方根变换和平方根变换的主要作用是把非线性数据结构变为线性数据结构，以适应某些统计方法的需要。

4.3 距离和相似系数的计算

研究变量或样本的亲疏程度的数量指标一般有两种，一种为距离，它是事物之间差异性的测度，将每一样本看成 n 维空间(n 个变量)的一个点，在这 n 维空间中定义距离，距离较近的点归为同一类，距离较远的点归为不同的类；另一种为相似系数，它是事物之间相似性的测度，性质越接近的样本，它们之间的相似系数越接近于1(或－1)。当聚类要素的数据处理工作完成以后，就要计算分类对象之间的距离或相似系数，并依据距离或相似系数的矩阵结构进行聚类。

4.3.1 距离的计算

如果我们把每一个分类对象的 n 个聚类要素看成 n 维空间的 n 个坐标轴，则每一个分类对象的 n 个要素所构成的 n 维数据向量就是 n 维空间中的一个点。这样，各分类对象之间的差异性就可以由它们所对应的 n 维空间中点之间的距离来度量。

假设有 m 个被聚类的对象，每一个被聚类对象都有 x_1，x_2，…，x_n 个要素构成。它们所对应的要素数据可用表4.6给出。第 i 个样本 x_i 为矩阵 $\boldsymbol{X}$

的第 i 行所描述，所以任何两个样本 x_K 与 x_L 之间的相似性可以通过矩阵 $\boldsymbol{X}$ 中第 K 行与第 L 行的相似度来刻画；任何两个变量 x_M 与 x_N 之间的相似性，可以通过矩阵第 M 列与第 N 列的相似度来刻画（K，L=1，2，…，m；M，N=1，2，…，n）。

表 4.6 聚类分析数据表

聚类对象	要素					
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}

其中，x_{ij} 表示第 i 个样本的第 j 个指标；d_{ij} 表示第 i 个样本和第 j 个样本之间的距离。

d_{ij} 应满足如下几个条件：

(1)非负性：$d_{ij} \geqslant 0 \quad (i, j=1, 2, \cdots, m)$；

(2)规范性：$d_{ij}=0 \quad (i=j=1, 2, \cdots, m)$；

(3)对称性：$d_{ij}=d_{ji} \quad (i, j=1, 2, \cdots, m)$；

(4)三角不等式：$d_{ij} \leqslant d_{ik}+d_{kj} \quad (i, j, k=1, 2, \cdots, m)$。

常用的距离有：

(1)绝对值距离(Kanhattan 度量或网格度量)

$$d_{ij}=\sum_{k=1}^{n} |x_{ik}-x_{jk}| \quad (i, j=1, 2, \cdots, m) \tag{4.5}$$

(2)欧氏距离(二阶 Minkowski 度量)

$$d_{ij}=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \quad (i, j=1, 2, \cdots, m) \tag{4.6}$$

欧氏距离是聚类分析中用得最广泛的距离。

(3)明科夫斯基(Minkowski)距离

$$d_{ij}=\left[\sum_{k=1}^{n} |x_{ik}-x_{jk}|^{p}\right]^{\frac{1}{p}} \quad (i, j=1, 2, \cdots, m) \tag{4.7}$$

式(4.7)中，$p \geqslant 1$。当 $p=1$ 时，它就是绝对值距离；当 $p=2$ 时，它就是欧氏距离。

(4)切比雪夫距离

当明科夫斯基距离 $p\to\infty$时，

$$d_{ij}(\infty)=\max|x_{ik}-x_{jk}|\quad(i,\ j=1,\ 2,\ \cdots,\ m)\tag{4.8}$$

(5)Canberra 度量(又称兰氏距离)

$$d_{ij}=\sum_{k=1}^{n}\frac{|x_{ik}-x_{jk}|}{x_{ik}+x_{jk}}\quad(i,\ j=1,\ 2,\ \cdots,\ m)\tag{4.9}$$

这是一个自身标准化的量，由于它对大的奇异值不敏感，所以特别适合高度偏倚的数据。

上述各种距离是假定变量之间相互独立，即在正交空间中讨论的距离。选择不同的距离，聚类结果会有所差异。在研究中，往往采用几种距离进行计算、对比，选择一种较为合理的距离进行聚类。

(6)马氏(P. C. Mahalanobis)距离

设 $\boldsymbol{S}$ 表示指标的协方差矩阵，即：$\boldsymbol{S}=(u_{ij})_{n\times n}$

其中，
$$u_{ij}=\frac{1}{m-1}\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)\quad(i,\ j=1,\ 2,\ \cdots,\ n)$$

$$\bar{x}_i=\frac{1}{m}\sum_{k=1}^{m}x_{ki},\qquad \bar{x}_j=\frac{1}{m}\sum_{k=1}^{m}x_{kj}$$

若 $\boldsymbol{S}^{-1}$存在，则两个样本之间的马氏距离为：

$$d_{ij}^2=(\boldsymbol{X}_i-\boldsymbol{X}_j)'\boldsymbol{S}^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j)\tag{4.10}$$

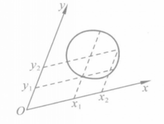
这里 $\boldsymbol{X}_i$ 为第 i 个样本的 n 个指标组成的向量，即原始资料阵的第 i 行向量；样本 $\boldsymbol{X}_j$ 类似。

马氏距离虽然可以排除变量之间相关性的干扰，并且不受量纲的影响，但是在聚类分析处理之前，如果用全部数据计算的均值和协方差阵来计算马氏距离，效果不是很好。比较合理的办法是用各个类的样本来计算各自的协方差矩阵，同一类样本的马氏距离应当用同一类的协方差矩阵来计算，而类的形成都要依赖于样本之间的距离，而样本之间合理的马氏距离又依赖于类，这就形成了一个恶性循环，因此在实际聚类分析处理中，马氏距离也不是理想的距离。为了克服变量间相关性的影响，可以引入斜交空间距离。

(7)斜交空间距离

由于多个变量之间存在着不同程度的相关关系，在这种情况下，用正交空间距离来计算样本间的距离，易产生变形，从而使聚类簇分类时的谱系结构发生变形。

图 4-1 表示在二维空间中，两个坐标轴在斜交和正交情况下，用欧氏距离计算所产生的变形，即斜交空间中的圆将在正交空间下变形为椭圆。



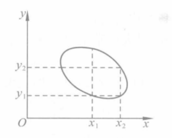


图 4-1　二维空间中，不同坐标轴系中用欧氏距离计算所产生的变形

在 n 维空间中，要使大量具有相关性的变量的谱系结构不发生变形，可采用斜交空间距离，其距离公式为：

$$d_{ij}=\left[\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}(x_{ik}-x_{jk})(x_{ik}-x_{jl})r_{kl}\right]^{\frac{1}{2}}\quad (i,\ j=1,\ 2,\ \cdots,\ m)\tag{4.11}$$

在数据标准化处理下，r_{kl} 为变量 k 和变量 l 之间的相关系数。

例 4.2　据表 4.3 中的数据，用式(4.5)～(4.7)计算 6 个监测站之间的绝对值距离、欧式距离和明科夫斯基距离。

解　6 个监测站之间的绝对值距离、欧式距离和明科夫斯基距离计算如下：

(1)当 $p=1$ 时，它就是绝对值距离。在 Matlab 环境下程序（说明：文字格式为程序中真实格式）为：

```
x=[-0.6186  -1.4259   0.1581  -0.6258  -0.4411;
   -0.3144  -0.8399  -0.1581  -0.3639  -0.6064;
   -0.3752  -0.1367  -1.1068  -0.3639  -0.4411;
   -1.3487  -0.3321  -1.2649   2.0812  -0.6064;
    1.2067   1.7384   1.5811   0.2474   2.2053;
    1.4501   0.3321   0.7906  -0.9751  -0.1103]
[m,n] =size(x);
a=zeros (m,m);
for i=1:m
    for j=1:m
        for k=1:n
            a(i,j)=a(i,j)+abs(x(i,k)-x(j,k));
end
end
end
a
```

$$
\boldsymbol{D}_1=(d_{ij})_{6\times 6}=\begin{pmatrix}
0.0000 & & & & & \\
1.6336 & 0.0000 & & & & \\
3.0594 & 1.8780 & 0.0000 & & & \\
6.7834 & 5.7582 & 4.2108 & 0.0000 & & \\
9.9322 & 9.2616 & 9.4026 & 11.4532 & 0.0000 & \\
5.1393 & 4.9925 & 5.1335 & 8.4067 & 5.9783 & 0.0000
\end{pmatrix}
$$

（2）当 $p=2$ 时，它就是欧氏距离。在 Matlab 环境下程序（说明：文字格式为程序中真实格式）为：

```
x=[-0.6186  -1.4259   0.1581  -0.6258  -0.4411;
   -0.3144  -0.8399  -0.1581  -0.3639  -0.6064;
   -0.3752  -0.1367  -1.1068  -0.3639  -0.4411;
   -1.3487  -0.3321  -1.2649   2.0812  -0.6064;
    1.2067   1.7384   1.5811   0.2474   2.2053;
    1.4501   0.3321   0.7906  -0.9751  -0.1103]
[m,n] =size(x);
b=zeros(m,m);
for i=1:m
    for j=1:m
        for k=1:n
            b(i,j)=b(i,j)+(x(i,k)-x(j,k))∧2;
end
end
end
sqrt(b)
```

$$
\boldsymbol{D}_2=(d_{ij})_{6\times 6}=\begin{pmatrix}
0.0000 & & & & & \\
0.7949 & 0.0000 & & & & \\
1.8412 & 1.1940 & 0.0000 & & & \\
3.6061 & 3.1059 & 2.6830 & 0.0000 & & \\
4.8099 & 4.5017 & 4.5410 & 5.2798 & 0.0000 & \\
2.8287 & 2.4509 & 2.7631 & 4.6525 & 3.0852 & 0.0000
\end{pmatrix}
$$

（3）当 $p>2$ 且为确定值时，可求得明科夫斯基距离矩阵。

4.3.2　相似系数的计算

聚类分析方法不仅可以用来对样本进行分类，而且可以对变量进行分类，在

对变量进行分类时，通常采用相似系数来表示变量之间的亲疏程度。两个变量越相似，它们的相似系数越大。在聚类分析中，总是把两个相似系数最大的变量首先归为一类。相似系数定义如下：

设 C_{ij} 表示变量 x_i 与 x_j 间的相似系数，则 C_{ij} 应满足如下关系：

(1) $C_{ij}=\pm 1 \Leftrightarrow x_i=ax_j \quad (a\in \mathbf{R},\ a\neq 0)$；

(2) $|C_{ij}|\leqslant 1 \quad (i,\ j=1,\ 2,\ \cdots,\ n)$；

(3) $C_{ij}=C_{ji} \quad (i,\ j=1,\ 2,\ \cdots,\ n)$。

当 $|C_{ij}|$ 越接近于1，则表示 x_i 与 x_j 关系越密切；$|C_{ij}|$ 越接近于0，则表示 x_i 与 x_j 关系越疏远。常见的相似系数是内积系数，主要包括：夹角余弦和相关系数，其计算公式如下。

①夹角余弦

图4-2中曲线 AB 和 CD 尽管长度不一，但形状相似，当长度不是主要矛盾时，应定义一种相似系数使 AB 和 CD 呈现出比较密切的关系，而夹角余弦适合这一要求。其定义为：

A C B D

图 4-2

将任何两个样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 看成 n 维空间的两个向量，这两个向量的夹角余弦用 $\cos\theta_{ij}$ 表示，则：

$$C_{ij}=\cos\theta_{ij}=\frac{\sum_{k=1}^{n}(x_{ik}x_{jk})}{\sqrt{\sum_{k=1}^{n}x_{ik}^2}\sqrt{\sum_{k=1}^{n}x_{jk}^2}} \quad (i,\ j=1,\ 2,\ \cdots,\ m) \tag{4.12}$$

在式(4.12)中，显然有：$-1\leqslant\cos\theta_{ij}\leqslant 1$。它是 $\boldsymbol{i}$ 和 $\boldsymbol{j}$ 两个指标向量在原点处的夹角 θ_{ij} 的余弦。当 $\cos\theta_{ij}=\pm 1$，说明两个样本 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 完全相似；$|\cos\theta_{ij}|$ 接近1，说明 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 相似密切；$\cos\theta_{ij}=0$，说明 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 完全不一样；$|\cos\theta_{ij}|$ 接近0，说明 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 差别大。因此，相似系数的数值范围为 $[-1,\ 1]$ 区间。

据表4.3中的数据，用夹角余弦公式式(4.12)计算，可得如下的夹角余弦矩阵：

$$\boldsymbol{C}_1=(\cos\theta_{ij})_{6\times 6}$$

$$=\begin{pmatrix} 1.0000 & 0.9284 & 0.2963 & -0.1752 & -0.6835 & -0.1730 \\ 0.9284 & 1.0000 & 0.5354 & -0.0135 & -0.8841 & -0.1949 \\ 0.2963 & 0.5354 & 1.0000 & 0.3653 & -0.7762 & -0.4160 \\ -0.1752 & -0.0135 & 0.3653 & 1.0000 & -0.3913 & -0.8597 \\ -0.6835 & -0.8841 & -0.7762 & -0.3913 & 1.0000 & 0.4598 \\ -0.1730 & -0.1949 & -0.4160 & -0.8597 & 0.4598 & 1.0000 \end{pmatrix}$$

②相关系数

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}}\quad (i,\ j=1,\ 2,\ \cdots,\ m)\quad(4.13)$$

在式(4.13)中，$\bar{x}_i$ 和 $\bar{x}_j$ 分别为聚类对象 i 和 j 各要素标准化数据的平均值。

据表4.3中的数据，用相关系数公式式(4.13)计算，可得如下的相关系数矩阵：

$$\boldsymbol{R}_2=(r_{ij})_{6\times6}$$

$$=\begin{pmatrix} 1.0000 & 0.8468 & -0.9128 & -0.4191 & 0.0116 & 0.1406 \\ 0.8468 & 1.0000 & -0.7612 & -0.2707 & -0.4196 & -0.2453 \\ -0.9128 & -0.7612 & 1.0000 & 0.4700 & -0.1116 & -0.2533 \\ -0.4191 & -0.2707 & 0.4700 & 1.0000 & -0.6564 & -0.8759 \\ 0.0116 & -0.4196 & -0.1116 & -0.6564 & 1.0000 & 0.3773 \\ 0.1406 & 0.2453 & -0.2533 & -0.8759 & 0.3773 & 1.0000 \end{pmatrix}$$

4.3.3　距离和相似系数选择原则

一般说来，同一批数据采用不同的相似性尺度，会得到不同的分类结果。产生不同结果的原因，主要是由于不同的指标所衡量的相似程度的物理意义不同，也就是说，不同指标代表了不同意义上的相似性。因此我们在进行数值分类时，应注意相似性尺度的选择，注意遵循下列基本选择原则：

(1)所选择的相似性尺度在实际应用中应有明确的意义，如在经济变量分析中，常用相关系数表示经济变量之间的亲疏程度。

(2)根据原始数据的性质，选择适当的变换方法。不同的变换方法涉及选用不同的相似系数，如标准化变换处理下，相关系数和夹角余弦一致；又如原始数据在进行聚类分析处理之前已经对变量的相关性作了处理，则通常可采用欧氏距离，而不必选用斜交空间距离。所选择的距离，还须和选用的聚类方法一致，如聚类方法选用离差平方和法时，距离只能采用欧氏距离。

(3)适当考虑计算工作量的大小，如对大样本的聚类问题，不适宜选择斜交空间距离，因采用该距离处理时，计算工作量太大。

距离的选择应根据研究对象，作具体分析。在多次进行聚类分析过程中，逐步总结经验，以选择合适的距离。初次进行聚类分析处理时，不妨用多选择几种计算距离的方法来进行聚类，作对比、分析，以确定合适的距离。

4.4 系统聚类分析常用方法

正如样本之间的距离可以有不同的定义方法一样，类与类之间的距离也有各种定义，例如可以定义类与类之间的距离为两类之间最近样本的距离，或者定义两类之间最远样本的距离，也可以定义两类重心之间的距离等。类与类之间用不同的方法定义距离，就产生了不同的系统聚类方法。

最短距离法、最远距离法、中间距离法、重心法、类平均法、可变类平均法、可变法和离差平方和法为常用的八种系统聚类方法。尽管系统聚类分析方法很多，但归类的步骤基本上一样，仅是类与类之间距离的定义方法有所不同，从而得到不同的计算距离的公式。这些公式在形式上不大一样，但最后可将它们统一为一个公式，为上机计算带来很大的方便，详见后文。本节重点介绍系统聚类方法。

系统聚类法是聚类分析诸多方法中应用较多的一个。它包含以下步骤：

(1)构造 m 个类，每个类只包含一个样本，记作 G_1，G_2，…，G_m；

(2)定义 m 个样本两两间的距离 $\{d_{ij}\}$，记作 $\boldsymbol{D}^{(0)}=(d_{ij}^{(0)})_{m\times m}$；

(3)合并距离最近的两类为一新类，记作 G_{m+1}类，并取消刚合并的那两类，得到 $m-1$ 类；

(4)计算新类与剩余各类的距离，若类的个数等于 1，转到步骤(5)，否则回到步骤(3)；

(5)画聚类图；

(6)确定临界值，决定类的个数和类的构成。

在某种意义上，最短距离法最优，类平均法和最远距离法次之。本节以一具体例题来解释最短距离法、最远距离法的具体计算过程。

在系统聚类法中，除了定义类间距离外，还要规定分类临界值，即聚类到某个“程度”时便停止，并非将所有的对象都归并为一大类。当类间距离大于给定的临界值时便停止聚类，由此得到若干个较少的类。

例 4.3 表 4.7 给出了某地区九个农业区的七项指标(徐建华，2006)①，经过极差标准化处理（见本章 4.2 节）后，如表 4.8 所示。

① http://218.24.233.167:8000/Resource/Book/Edu/JXCKS/TS090038/0007_ts090038.htm

表 4.7 某地区九个农业区的七项指标数据

区代号	人均耕地 X_1 /(hm^2·人$^{-1}$)	劳动耕地 X_2 /(hm^2·个$^{-1}$)	水田比重 X_3 /%	复种指数 X_4 /%	粮食亩产 X_5 /(kg·hm^{-2})	人均粮食 X_6 /(kg·人$^{-1}$)	稻谷占粮食比重 X_7 /%
G_1	0.294	1.093	5.63	113.6	4 510.5	1 036.4	12.20
G_2	0.315	0.971	0.39	95.1	2 773.5	683.7	0.85
G_3	0.123	0.316	5.28	148.5	6 934.5	611.1	6.49
G_4	0.179	0.527	0.39	111.0	4 458.0	632.6	0.92
G_5	0.081	0.212	72.04	217.8	12 249.0	791.1	80.38
G_6	0.082	0.211	43.78	179.6	8 973.0	636.5	48.17
G_7	0.075	0.181	65.15	194.7	10 689.0	634.3	80.17
G_8	0.293	0.666	5.35	94.9	3 679.5	771.7	7.80
G_9	0.167	0.414	2.90	94.8	4 231.5	574.6	1.17

表 4.8 极差标准化后的数据

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
G_1	0.912 5	1.000 0	0.073 1	0.152 8	0.183 3	1.000 0	0.142 7
G_2	1.000 0	0.866 2	0.000 0	0.002 4	0.000 0	0.236 2	0.000 0
G_3	0.200 0	0.148 0	0.068 2	0.436 6	0.439 1	0.079 0	0.070 9
G_4	0.433 3	0.379 4	0.000 0	0.131 7	0.177 8	0.125 6	0.000 9
G_5	0.025 0	0.034 0	1.000 0	1.000 0	1.000 0	0.468 8	1.000 0
G_6	0.029 2	0.032 9	0.605 6	0.689 4	0.654 3	0.134 0	0.595 0
G_7	0.000 0	0.000 0	0.903 8	0.812 2	0.835 4	0.129 3	0.997 4
G_8	0.908 3	0.531 8	0.069 2	0.000 8	0.095 6	0.426 8	0.087 4
G_9	0.383 3	0.255 5	0.035 0	0.000 0	0.153 9	0.000 0	0.004 0

4.4.1 最短距离系统聚类法原理

原理：最短距离聚类法，是在原来的 $m\times m$ 距离矩阵的非对角元素中找出最小值 $d_{p,q}$，把分类对象 G_p 和 G_q 归并为一新类 G_r，然后按计算公式：

$$d_{r,k}=\min\{d_{p,k},\ d_{q,k}\}\quad (k\neq p,\ q)\tag{4.14}$$

计算原来各类与新类之间的距离，这样就得到一个新的 $(m-1)$ 阶的距离矩阵；再从新的距离矩阵中选出最小者 $d_{i,j}$，把 G_i 和 G_j 归并成新类；再计算各类与新类的距离，这样一直计算下去，直至各分类对象被归为一类为止。

例 4.4 已知九个农业区之间的绝对值距离矩阵，使用最短距离聚类法作聚类分析。

$$\boldsymbol{D}_3=(d_{i,j})_{(9\times 9)}$$

$$=\begin{pmatrix}
0.0000 \\
1.5346 & 0.0000 \\
3.1017 & 2.6879 & 0.0000 \\
2.2158 & 1.4721 & 1.2158 & 0.0000 \\
5.8327 & 6.0374 & 3.6639 & 4.7866 & 0.0000 \\
4.7087 & 4.4482 & 1.8704 & 2.9930 & 1.7958 & 0.0000 \\
5.7800 & 5.5195 & 2.9321 & 4.0548 & 0.8498 & 1.0713 & 0.0000 \\
1.3445 & 0.8705 & 2.2366 & 1.2974 & 5.1701 & 3.9621 & 5.0334 & 0.0000 \\
2.6328 & 1.6590 & 1.1918 & 0.4933 & 4.8557 & 3.0621 & 4.1239 & 1.4048 & 0.0000
\end{pmatrix}$$

根据上面的矩阵，用最短距离聚类法作聚类分析：

(1)在 9×9 阶距离矩阵 $\boldsymbol{D}_3$ 中，非对角元素中最小者是 $d_{9,4}=0.4933$，故首先将第 4 区与第 9 区并为一类，记为 G_{10}，即 $G_{10}=\{G_4,\ G_9\}$。按式(4.14)分别计算 G_1，G_2，G_3，G_5，G_6，G_7，G_8 与 G_{10}之间的距离，得：

$$d_{1,10}=\min\{d_{1,4},\ d_{1,9}\}=\min\{2.2158,\ 2.6328\}=2.2158$$
$$d_{2,10}=\min\{d_{2,4},\ d_{2,9}\}=\min\{1.4721,\ 1.6590\}=1.4721$$
$$d_{3,10}=\min\{d_{3,4},\ d_{3,9}\}=\min\{1.2158,\ 1.1918\}=1.1918$$
$$d_{5,10}=\min\{d_{5,4},\ d_{5,9}\}=\min\{4.7866,\ 4.8557\}=4.7866$$
$$d_{6,10}=\min\{d_{6,4},\ d_{6,9}\}=\min\{2.9930,\ 3.0621\}=2.9930$$
$$d_{7,10}=\min\{d_{7,4},\ d_{7,9}\}=\min\{4.0548,\ 4.1239\}=4.0548$$
$$d_{8,10}=\min\{d_{8,4},\ d_{8,9}\}=\min\{1.2974,\ 1.4048\}=1.2974$$

这样就得到 G_1，G_2，G_3，G_5，G_6，G_7，G_8，G_{10}上的一个新的 8×8 阶距离矩阵：

	G_1	G_2	G_3	G_5	G_6	G_7	G_8	G_{10}
G_1	0.000 0							
G_2	1.534 6	0.000 0						
G_3	3.101 7	2.687 9	0.000 0					
G_5	5.832 7	6.037 4	3.663 9	0.000 0				
G_6	4.708 7	4.448 2	1.870 4	1.795 8	0.000 0			
G_7	5.780 0	5.519 5	2.932 1	0.849 8	1.071 3	0.000 0		
G_8	1.344 5	0.870 5	2.236 6	5.170 1	3.962 1	5.033 4	0.000 0	
G_{10}	2.215 8	1.472 1	1.191 8	4.786 6	2.993 0	4.054 8	1.297 4	0.000 0

(2)在上一步骤中所得到的8×8阶距离矩阵中，非对角元素中最小者为$d_{5,7}=0.849\ 8$，故将G_5与G_7归并为一类，记为G_{11}，即$G_{11}=\{G_5, G_7\}$。

按此方法类推可将所有分类对象归类。

综合上述聚类过程，可以作出最短距离聚类谱系图。

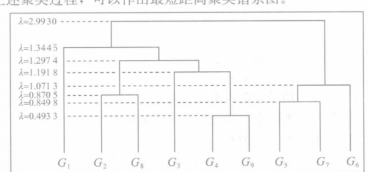


图4-3 最短距离聚类谱系图

最后，决定类的个数与类。如果在图4-3距离为1.297 4处切一刀，即得到九个农业区的三大类。它们分别是$\{G_1\}$，$\{G_2, G_3, G_4, G_8, G_9\}$及$\{G_5, G_6, G_7\}$。在实际问题中有时给出一个阈值$T$，要求类与类之间的距离小于$T$，因此有些样本可能归不了类。

4.4.2 最远距离聚类法原理

原理：最远距离聚类法，是在原来的$m\times m$距离矩阵的非对角元素中找出最小值$d_{p,q}$，把分类对象G_p和G_q归并为一新类G_r，然后按计算公式：

$$d_{k,r}=\max\{d_{k,p}, d_{k,q}\} \quad (k\neq p, q) \tag{4.15}$$

计算原来各类与新类之间的距离，这样就得到一个新的$(m-1)$阶的距离矩阵；再从新的距离矩阵中选出最小者 $d_{i,j}$，把 G_i 和 G_j 归并成新类；再计算各类与新类的距离，这样一直计算下去，直至各分类对象被归为一类为止。

最远距离聚类法与最短距离聚类法的区别在于计算原来的类与新类距离时，采用的公式不同，而其并类步骤完全一样。

例 4.5 已知九个农业区之间的绝对值距离矩阵，使用最远距离聚类法作聚类分析。

$$\boldsymbol{D}_3=(d_{i,j})_{(9\times9)}$$

$$=\begin{pmatrix}
0.0000 \\
1.5346 & 0.0000 \\
3.1017 & 2.6879 & 0.0000 \\
2.2158 & 1.4721 & 1.2158 & 0.0000 \\
5.8327 & 6.0374 & 3.6639 & 4.7866 & 0.0000 \\
4.7087 & 4.4482 & 1.8704 & 2.9930 & 1.7958 & 0.0000 \\
5.7800 & 5.5195 & 2.9321 & 4.0548 & 0.8498 & 1.0713 & 0.0000 \\
1.3445 & 0.8705 & 2.2366 & 1.2974 & 5.1701 & 3.9621 & 5.0334 & 0.0000 \\
2.6328 & 1.6590 & 1.1918 & 0.4933 & 4.8557 & 3.0621 & 4.1239 & 1.4048 & 0.0000
\end{pmatrix}$$

根据上面的矩阵，用最远距离聚类法聚类：

(1)在 9×9 阶距离矩阵中，非对角元素中最小者是 $d_{4,9}=0.4933$，故首先将第 4 区与第 9 区并为一类，记为 G_{10}，即 $G_{10}=\{G_4, G_9\}$。按式(4.15)分别计算 G_1，G_2，G_3，G_5，G_6，G_7，G_8 与 G_{10}之间的距离，得：

$$d_{1,10}=\max\{d_{1,4}, d_{1,9}\}=\max\{2.2158, 2.6328\}=2.6328$$
$$d_{2,10}=\max\{d_{2,4}, d_{2,9}\}=\max\{1.4721, 1.6590\}=1.6590$$
$$d_{3,10}=\max\{d_{3,4}, d_{3,9}\}=\max\{1.2158, 1.1918\}=1.2158$$
$$d_{5,10}=\max\{d_{5,4}, d_{5,9}\}=\max\{4.7866, 4.8557\}=4.8557$$
$$d_{6,10}=\max\{d_{6,4}, d_{6,9}\}=\max\{2.9930, 3.0621\}=3.0621$$
$$d_{7,10}=\max\{d_{7,4}, d_{7,9}\}=\max\{4.0548, 4.1239\}=4.1239$$
$$d_{8,10}=\max\{d_{8,4}, d_{8,9}\}=\max\{1.2974, 1.4048\}=1.4048$$

这样就得到 G_1，G_2，G_3，G_5，G_6，G_7，G_8，G_{10}上的一个新的 8×8 阶距离矩阵：

	G_1	G_2	G_3	G_5	G_6	G_7	G_8	G_{10}
G_1	0.000 0							
G_2	1.534 6	0.000 0						
G_3	3.101 7	2.687 9	0.000 0					
G_5	5.832 7	6.037 4	3.663 9	0.000 0				
G_6	4.708 7	4.448 2	1.870 4	1.795 8	0.000 0			
G_7	5.780 0	5.519 5	2.932 1	0.849 8	1.071 3	0.000 0		
G_8	1.344 5	0.870 5	2.236 6	5.170 1	3.962 1	5.033 4	0.000 0	
G_{10}	2.632 8	1.658 9	1.215 8	4.855 7	3.062 1	4.123 9	1.404 8	0.000 0

(2)在第一步得到的 8×8 阶距离矩阵中，非对角线元素中最小者为 $d_{5,7}=0.849\ 8$，故将 G_5 与 G_7 归并为一类，记为 G_{11}，即 $G_{11}=\{G_5, G_7\}$。

按式（4.15）计算，依次归类。

综合上述聚类过程，可以作出最远距离聚类谱系图。

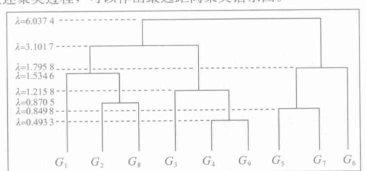


图 4-4　最远距离聚类谱系图

应注意，最短距离法也可用于对指标(变量)的分类，分类时可以用距离，也可以用相似系数。但用相似系数时应找到最大的元素并类，计算新类与其他类的距离应使用最远距离公式(4.15)。

以上仅为两种系统聚类方法，其余的系统聚类方法，读者可参看有关文献。

4.4.3　系统聚类法公式的统一

系统聚类法通常有八种方法（表 4.9），这些方法的分类原则和过程基本是一致的，所不同的是类与类之间的距离有不同的定义。能否将它们统一起来呢？关键在于八种类型之间的距离的定义能否统一。1969 年，Wishart 将八种不同的距离计算公式统一为如下递推公式：

$$d_{kr}^2=\alpha_p d_{kp}^2+\alpha_q d_{kq}^2+\beta d_{pq}^2+\gamma\mid d_{kp}^2-d_{kq}^2\mid \tag{4.16}$$

应用这个递推公式的前提是：设类 G_p 和类 G_q 合并为新类 G_r，当计算新类 G_r 和 $G_k(k\neq p,\ q)$之间的距离 d_{kr}^2 就用公式(4.16)。式中，参数 α_p，α_q，β，γ 取不同的值时，就形成了不同的聚类方法(表 4.9)。

下面是最短距离聚类法和最远距离聚类法公式的统一：

(1)最短距离聚类法具有空间压缩性，而最远距离聚类法具有空间扩张性(图 4-5)。最短距离为 $d_{AB}=d_{a_1b_1}$，最远距离为 $d_{AB}=d_{a_2b_2}$。

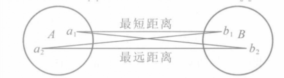


图 4-5　最短距离聚类法和最远距离聚类法之间的关系图

(2)最短距离聚类法和最远距离聚类法关于类之间的距离计算可以用统一的式子表示：

$$d_{kr}^2=\alpha_p d_{kp}^2+\alpha_q d_{kq}^2+\gamma\mid d_{kp}^2-d_{kq}^2\mid \tag{4.17}$$

当 $\gamma=-1/2$ 时，就是最短距离聚类法计算类间距离的公式；当 $\gamma=1/2$ 时，就是最远距离聚类法计算类间距离的公式。各种系统聚类法类间距离关系，见表 4.9。

表 4.9　八种不同系统聚类方法计算类间距离的统一表达式(胡永宏等，2000)

方法名称	参数				距离矩阵要求	空间性质
	α_p	α_q	β	γ		
最短距离法	1/2	1/2	0	−1/2	各种距离	压缩
最远距离法	1/2	1/2	0	1/2	各种距离	扩张
中间距离法	1/2	1/2	[−1/4, 0]	0	欧氏距离	保持
重心法	$\frac{n_p}{n_p+n_q}$	$\frac{n_q}{n_p+n_q}$	$\frac{-n_p\times n_q}{(n_p+n_q)^2}$	0	欧氏距离	保持
类平均法	$\frac{n_p}{n_p+n_q}$	$\frac{n_q}{n_p+n_q}$	0	0	各种距离	保持
离差平方和法	$\frac{n_k+n_p}{n_k+n_r}$	$\frac{n_k+n_q}{n_k+n_r}$	$\frac{-n_k}{n_k+n_r}$	0	欧氏距离	压缩
可变类平均法	$\frac{(1-\beta)n_p}{n_r}$	$\frac{(1-\beta)n_q}{n_r}$	<1	0	各种距离	不定
可变法	$\frac{(1-\beta)}{2}$	$\frac{(1-\beta)}{2}$	<1	0	各种距离	扩张

注：n_p，n_q，n_r，n_k 分别是 G_p，G_q，G_r，G_k 的样本数目。

对于八种系统聚类方法，使用的情况和优劣各不相同。不同的聚类方法各有优点，但也存在其欠缺的一面。最短距离法分类最为简单，应用也较多，但当两类合并后与其他类的距离是所有距离中的最小者，从而缩小了新合并类与其他类的距离，产生空间收缩，因而其灵敏度比较低。最远距离法正好与最短距离法相反，两类合并后产生空间扩张。中间距离法既不是采用两类间最近距离，也不是采用最远距离，而是采用介于最近与最远之间的距离。重心法有较好的代表性，但计算繁琐，而且没有充分利用各样本的信息。类平均法被认为是较好的方法之一，但在递推公式中没有反映类 G_p 和类 G_q 的距离，这是其不足的一面。在可变法和可变类平均法中加进了 β 因子，但就具体的问题确定 β 值不是易事。离差平方和法是八种方法中最有统计特点的一种方法，它基于方差分析的思想，所以如果分类得当，同类样本之间的离差平方和应当较小，而类间的离差平方和应当较大。

4.5 环境应用

例 4.6　如图 4－6 所示，以长江流域为例，选取 20 个监测站点，对其水质污染水平进行类型划分及差异性程度分析。



图 4－6　长江流域监测站图

(1)聚类指标选择

选取如下 5 项指标作为对长江流域 20 个监测站点水质污染水平进行聚类分析的基础指标：

①DO——溶解氧，反映水体自净能力大小；

②BOD_5——生化需氧量，反映水体中能被生物降解的有机需氧污染物质含量；

③高锰酸盐指数——反映水体中部分有机耗氧物质含量的指标；

④NH_3-N——氨氮，反映水体受含氮有机物污染程度的指标；

⑤挥发酚——反映水中酚类有毒物质的含量的指标。

表 4.10　　1997 年 1 月和 7 月各站点水环境监测指标实测值　单位：mg/L

各站点	指标 1月份					指标 7月份				
	DO	高锰酸盐指数	BOD_5	NH_3-N	挥发酚	DO	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
攀枝花	10.4	2.3	3.8	0.16	0.002	8.9	5.0	1.8	0.21	0.000
望江楼	0.8	8.2	6.6	3.91	0.031	0.8	8.2	6.6	3.91	0.031
高　场	10.0	2.7	1.4	0.22	0.004	8.0	3.1	1.6	0.26	0.000
朱　沱	10.4	1.6	1.1	0.37	0.000	7.2	1.4	1.5	0.38	0.000
寸　滩	10.5	1.5	1.8	0.45	0.000	6.7	1.8	1.5	0.46	0.000
贵　阳	3.6	23.7	71.3	13.05	0.020	6.0	2.2	1.9	1.18	0.000
张家界	10.2	2.1	1.4	0.07	0.000	8.4	1.4	1.0	0.37	0.000
吉　首	10.2	2.8	5.0	0.75	0.003	7.4	2.4	1.1	0.47	0.000
芷　江	9.2	1.7	1.3	0.06	0.000	7.0	2.5	0.4	0.11	0.000
坝　上	10.3	3.6	1.8	0.00	0.000	6.6	2.1	0.8	0.00	0.000
津　市	10.4	1.8	0.7	0.35	0.000	8.5	2.2	0.9	0.22	0.000
石　门	10.4	1.7	0.8	0.36	0.008	7.1	2.4	1.2	0.58	0.000
益　阳	10.3	1.7	0.6	0.49	0.000	6.9	1.6	1.1	0.24	0.000
湘　潭	9.6	2.4	0.4	1.16	0.007	5.3	2.7	0.3	0.18	0.000
株　洲	5.4	2.8	1.1	0.28	0.000	6.2	2.5	1.1	0.10	0.006
衡　阳	6.8	2.3	2.9	0.49	0.007	7.1	3.3	1.1	0.20	0.000
长　沙	10.3	2.7	1.7	1.35	0.000	6.8	2.0	0.4	0.50	0.002
吉　安	9.8	2.0	2.8	0.00	0.000	5.6	2.4	1.4	0.17	0.000
中　山	12.8	2.1	2.5	0.10	0.000	7.3	0.8	1.2	0.00	0.027
宣　城	14.6	1.0	3.6	0.00	0.000	8.1	1.2	1.7	0.08	0.000

(2)系统聚类计算

①用标准差标准化方法（见本章4.2节）对5项指标的原始数据进行处理，见表4.11。

②采用欧氏距离（见本章4.3.1）测度20个监测站点之间的样本间距离，见$\boldsymbol{D}_1$，$\boldsymbol{D}_7$(分别表示1月份和7月份)。

③选用最远距离法计算类间的距离，并对样本进行归类，见图4-7。

表4.11　　标准差标准化法处理后的数据

各站点	指标 1月份					指标 7月份				
	DO	高锰酸盐指数	BOD_5	NH_3-N	挥发酚	DO	高锰酸盐指数	BOD_5	NH_3-N	挥发酚
1	0.368 4	−0.255 0	−0.120 8	−0.358 0	−0.269 3	1.272 5	1.564 8	0.292 5	−0.327 4	−0.379 5
2	−2.847 0	0.963 2	0.064 0	0.956 9	3.450 1	−3.624 0	3.617 0	4.087 1	4.142 9	3.185 6
3	0.234 5	−0.172 4	−0.279 3	−0.337 0	−0.012 8	0.728 4	0.346 3	0.134 4	−0.267 0	−0.379 5
4	0.368 4	−0.399 5	−0.299 1	−0.284 4	−0.525 9	0.244 8	−0.743 9	0.055 3	−0.122 0	−0.379 5
5	0.401 9	−0.420 2	−0.252 9	−0.256 3	−0.525 9	−0.057 4	−0.487 4	0.055 3	−0.025 4	−0.379 5
6	−1.909 1	4.163 7	4.336 2	4.161 8	2.039 3	−0.480 6	−0.230 9	0.371 6	0.844 5	−0.379 5
7	0.301 4	−0.296 3	−0.279 3	−0.389 6	−0.525 9	0.970 2	−0.743 9	−0.339 9	−0.134 1	−0.379 5
8	0.301 4	−0.151 8	−0.041 6	−0.151 1	−0.141 1	0.365 7	−0.102 6	−0.260 9	−0.013 3	−0.379 5
9	−0.033 5	−0.378 9	−0.285 9	−0.393 1	−0.525 9	0.123 9	−0.038 5	−0.814 3	−0.448 2	−0.379 5
10	0.334 9	0.013 4	−0.252 9	−0.414 1	−0.525 9	−0.117 9	−0.295 0	−0.498 0	−0.581 1	−0.379 5
11	0.368 4	−0.358 2	−0.325 5	−0.291 4	−0.525 9	1.030 7	−0.230 9	−0.419 0	−0.315 3	−0.379 5
12	0.368 4	−0.378 9	−0.318 9	−0.287 9	0.500 2	0.184 4	−0.102 6	−0.181 8	0.119 6	−0.379 5
13	0.334 9	−0.378 9	−0.332 1	−0.242 3	−0.525 9	0.063 5	−0.615 7	−0.260 9	−0.291 2	−0.379 5
14	0.100 5	−0.234 4	−0.345 3	−0.007 4	0.371 9	−0.903 7	0.089 8	−0.893 3	−0.363 7	−0.379 5
15	−1.306 3	−0.151 8	−0.299 1	−0.315 9	−0.525 9	−0.359 7	−0.038 5	−0.260 9	−0.460 3	0.310 5
16	−0.837 3	−0.255 0	−0.180 3	−0.242 3	0.371 9	0.184 4	0.474 6	−0.260 9	−0.339 5	−0.379 5
17	0.334 9	−0.172 4	−0.259 5	0.059 3	−0.525 9	0.003 0	−0.359 1	−0.814 3	0.023 0	−0.149 5
18	0.167 5	−0.316 9	−0.186 9	−0.414 1	−0.525 9	−0.722 4	−0.102 6	−0.023 7	−0.375 8	−0.379 5
19	1.172 3	−0.296 3	−0.206 7	−0.379 0	−0.525 9	0.305 3	−1.128 7	−0.181 8	−0.581 1	2.725 6
20	1.775 2	−0.523 4	−0.134 0	−0.414 1	−0.525 9	0.788 9	−0.872 2	0.213 4	−0.484 5	−0.379 5

$$D_1=(d_{ij})_{20\times 20}=\begin{pmatrix}
0.000\,0 \\
5.236\,4 & 0.000\,0 \\
0.340\,7 & 4.956\,7 & 0.000\,0 \\
0.352\,0 & 5.447\,7 & 0.579\,6 & 0.000\,0 \\
0.349\,4 & 5.463\,5 & 0.599\,9 & 0.066\,9 & 0.000\,0 \\
8.386\,6 & 6.452\,6 & 8.315\,6 & 8.593\,3 & 8.573\,9 & 0.000\,0 \\
0.313\,3 & 5.407\,2 & 0.534\,7 & 0.163\,1 & 0.209\,6 & 8.566\,0 & 0.000\,0 \\
0.284\,0 & 5.029\,0 & 0.335\,3 & 0.545\,9 & 0.534\,7 & 8.125\,9 & 0.531\,4 & 0.000\,0 \\
0.520\,8 & 5.241\,2 & 0.617\,2 & 0.417\,1 & 0.459\,4 & 8.534\,9 & 0.345\,0 & 0.655\,8 & 0.000\,0 \\
0.399\,5 & 5.368\,0 & 0.560\,8 & 0.436\,5 & 0.466\,3 & 8.416\,6 & 0.313\,6 & 0.538\,8 & 0.539\,6 & 0.000\,0 \\
0.350\,5 & 5.440\,9 & 0.565\,6 & 0.049\,5 & 0.107\,1 & 8.589\,4 & 0.141\,8 & 0.543\,5 & 0.417\,0 & 0.399\,4 & 0.000\,0 \\
0.807\,2 & 4.747\,4 & 0.572\,5 & 1.026\,5 & 1.030\,1 & 8.346\,3 & 1.037\,4 & 0.750\,3 & 1.107\,5 & 1.108\,2 & 1.026\,3 & 0.000\,0 \\
0.374\,6 & 5.415\,7 & 0.572\,5 & 0.066\,4 & 0.112\,5 & 8.569\,7 & 0.180\,1 & 0.541\,7 & 0.400\,7 & 0.435\,5 & 0.063\,3 & 1.027\,7 & 0.000\,0 \\
0.810\,3 & 4.549\,1 & 0.531\,8 & 0.991\,9 & 1.000\,9 & 8.090\,8 & 1.000\,3 & 0.650\,6 & 0.998\,6 & 1.047\,1 & 0.987\,0 & 0.434\,2 & 0.968\,1 & 0.000\,0 \\
1.707\,2 & 4.601\,9 & 1.624\,4 & 1.693\,2 & 1.730\,8 & 8.191\,6 & 1.616\,0 & 1.681\,1 & 1.295\,3 & 1.653\,1 & 1.687\,8 & 1.977\,4 & 1.658\,8 & 1.699\,8 & 0.000\,0 \\
1.371\,8 & 4.061\,5 & 1.149\,9 & 1.515\,4 & 1.540\,9 & 7.952\,9 & 1.461\,5 & 1.264\,1 & 1.225\,3 & 1.512\,3 & 1.514\,6 & 1.227\,5 & 1.489\,5 & 0.981\,0 & 1.027\,7 & 0.000\,0 \\
0.516\,9 & 5.304\,1 & 0.656\,4 & 0.415\,2 & 0.406\,9 & 8.268\,5 & 0.467\,3 & 0.491\,3 & 0.619\,5 & 0.508\,6 & 0.403\,7 & 1.104\,9 & 0.372\,7 & 0.936\,3 & 1.684\,1 & 1.511\,3 & 0.000\,0 \\
0.342\,9 & 5.336\,4 & 0.550\,6 & 0.276\,8 & 0.308\,0 & 8.506\,8 & 0.165\,8 & 0.532\,5 & 0.233\,4 & 0.376\,1 & 0.276\,3 & 1.063\,2 & 0.287\,2 & 1.003\,9 & 1.490\,5 & 1.359\,8 & 0.527\,5 & 0.000\,0 \\
0.849\,5 & 5.950\,4 & 1.079\,4 & 0.821\,2 & 0.791\,2 & 8.787\,7 & 0.874\,0 & 1.003\,3 & 1.211\,3 & 0.894\,7 & 0.819\,7 & 1.314\,1 & 0.861\,7 & 1.454\,6 & 2.485\,3 & 2.205\,8 & 0.954\,7 & 1.005\,8 & 0.000\,0 \\
1.456\,1 & 6.426\,7 & 1.669\,5 & 1.427\,8 & 1.391\,3 & 9.112\,9 & 1.498\,5 & 1.592\,5 & 1.820\,9 & 1.541\,7 & 1.434\,6 & 1.761\,5 & 1.471\,1 & 1.975\,9 & 3.109\,8 & 2.781\,2 & 1.561\,3 & 1.621\,8 & 0.649\,3 & 0.000\,0
\end{pmatrix}$$

$$D_7=(d_{ij})_{20\times 20}=\begin{pmatrix}
0.000\,0 \\
8.676\,4 & 0.000\,0 \\
1.345\,2 & 8.799\,0 & 0.000\,0 \\
2.546\,5 & 9.007\,8 & 1.204\,0 & 0.000\,0 \\
2.475\,4 & 8.712\,5 & 1.173\,5 & 0.408\,0 & 0.000\,0 \\
2.770\,8 & 7.879\,2 & 1.756\,9 & 1.350\,4 & 1.049\,6 & 0.000\,0 \\
2.420\,5 & 9.525\,0 & 1.220\,5 & 0.826\,2 & 1.135\,7 & 1.957\,5 & 0.000\,0 \\
2.001\,9 & 8.868\,0 & 0.744\,1 & 0.733\,3 & 0.653\,6 & 1.367\,0 & 0.893\,0 & 0.000\,0 \\
2.264\,8 & 9.231\,5 & 1.202\,6 & 1.172\,5 & 1.081\,4 & 1.865\,4 & 1.240\,0 & 0.747\,0 & 0.000\,0 \\
2.466\,0 & 9.145\,8 & 1.275\,2 & 0.921\,9 & 0.809\,7 & 1.710\,0 & 1.269\,0 & 0.805\,9 & 0.491\,9 & 0.000\,0 \\
1.946\,6 & 9.453\,2 & 0.856\,2 & 1.069\,2 & 1.248\,5 & 2.062\,6 & 0.553\,1 & 0.758\,2 & 1.016\,5 & 1.183\,3 & 0.000\,0 \\
2.095\,0 & 8.686\,8 & 0.864\,2 & 0.727\,7 & 0.532\,7 & 1.136\,0 & 1.057\,4 & 0.238\,3 & 0.854\,5 & 0.848\,2 & 0.989\,0 & 0.000\,0 \\
2.554\,2 & 9.098\,9 & 1.234\,7 & 0.421\,8 & 0.449\,1 & 1.460\,8 & 0.932\,4 & 0.657\,1 & 0.817\,1 & 0.525\,4 & 1.053\,1 & 0.673\,0 & 0.000\,0 \\
2.884\,3 & 8.812\,8 & 1.948\,1 & 1.724\,1 & 1.436\,5 & 1.828\,0 & 2.136\,7 & 1.473\,5 & 1.042\,0 & 0.984\,4 & 2.017\,9 & 1.400\,3 & 1.355\,9 & 0.000\,0 \\
2.456\,6 & 8.507\,6 & 1.414\,8 & 1.246\,4 & 1.028\,6 & 1.621\,8 & 1.689\,7 & 1.098\,3 & 1.008\,2 & 0.819\,2 & 1.578\,7 & 1.057\,7 & 1.008\,4 & 1.094\,4 & 0.000\,0 \\
1.636\,7 & 8.722\,7 & 0.688\,4 & 1.278\,9 & 1.087\,5 & 1.655\,9 & 1.466\,5 & 0.687\,3 & 0.764\,9 & 0.893\,5 & 1.113\,3 & 0.741\,7 & 1.098\,0 & 1.316\,3 & 1.024\,7 & 0.000\,0 \\
2.591\,1 & 9.004\,7 & 1.435\,6 & 1.018\,2 & 0.911\,9 & 1.544\,2 & 1.177\,4 & 0.746\,9 & 0.626\,4 & 0.732\,5 & 1.181\,6 & 0.749\,0 & 0.726\,2 & 1.110\,1 & 0.992\,9 & 1.103\,9 & 0.000\,0 \\
2.619\,6 & 8.501\,9 & 1.530\,7 & 1.190\,5 & 0.848\,1 & 1.311\,6 & 1.853\,3 & 1.171\,2 & 1.162\,2 & 0.818\,3 & 1.802\,7 & 1.045\,3 & 0.971\,8 & 0.909\,0 & 0.821\,7 & 1.101\,4 & 1.195\,4 & 0.000\,0 \\
4.256\,9 & 8.872\,0 & 3.492\,1 & 3.171\,8 & 3.248\,0 & 3.661\,1 & 3.233\,7 & 3.320\,7 & 3.358\,7 & 3.258\,2 & 3.331\,8 & 3.346\,7 & 3.170\,7 & 3.625\,1 & 2.735\,7 & 3.505\,9 & 3.116\,9 & 3.437\,7 & 0.000\,0 \\
2.490\,7 & 9.421\,0 & 1.241\,8 & 0.684\,8 & 1.048\,8 & 1.953\,0 & 0.691\,6 & 1.103\,8 & 1.481\,5 & 1.292\,6 & 0.947\,8 & 1.216\,1 & 0.924\,3 & 2.242\,7 & 1.648\,0 & 1.557\,3 & 1.499\,2 & 1.715\,9 & 3.179\,1 & 0.000\,0
\end{pmatrix}$$

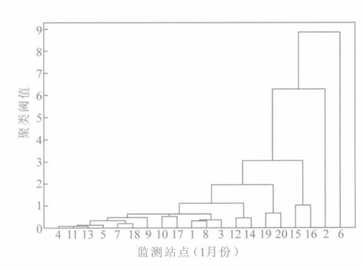


图4-7　1997年1月份20个水质监测站点最远距离聚类图

由图4-7及表4.11可知，1997年1月份20个水质监测站点水质污染水平差异性程度分为五类较合适，即{1，3，4，5，7，8，9，10，11，12，13，14，17，18}为一类；{19，20}为一类；{15，16}为一类；2，6号两监测站点不能归类，故各自为一类。其中2，6号监测站点水质恶劣，按照GB 3838—2002地表水环境质量标准，五项监测指标中只有一项为Ⅳ类水质，其余均属Ⅴ类水，2，6号监测站点1月份整体水质均为Ⅴ类水。

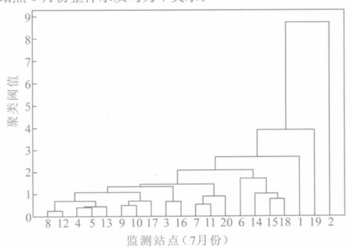


图4-8　1997年7月份20个水质监测站点最远距离聚类图

由图4-8及表4.11可知，1997年7月份20个水质监测站点水质污染水平差异性程度分为五类较合适，即{3，4，5，7，8，9，10，11，12，13，16，17，20}为一类；{6，14，15，18}为一类；1，2，19号三个监测站点不能归类。按照GB 3838—2002地表水环境质量标准，1号站点水质基本属Ⅲ类水质；2号站点7月份与1月份水质相同，因为该地区汛期水量虽增大，但排污量也相应增大；19号监测站点水质较好，五项监测指标中只有挥发酚为Ⅳ类水质，其余均属Ⅰ类水质。

【思考题4】

1. 给出聚类分析数据表4.12的明科夫斯基距离矩阵和相似系数矩阵。

要求：(1) 采用标准差标准法对聚类数据进行标准化，并求 $p=4$ 时的明科夫斯基距离矩阵；

(2) 采用极差标准化法对聚类数据进行标准化，求聚类对象之间的夹角余弦矩阵。

表4.12　聚类数据

聚类对象	要素					
1	0.046	0.087	0.031	0.038	0.008	0.022 0
2	0.049	0.055	0.100	0.110	0.022	0.007 3
3	0.038	0.130	0.079	0.170	0.058	0.043 0
4	0.034	0.084	0.058	0.160	0.200	0.029 0
5	0.084	0.066	0.029	0.320	0.012	0.041 0
6	0.064	0.072	0.100	0.210	0.028	1.380 0
7	0.048	0.089	0.062	0.260	0.038	0.036 0
8	0.059	0.055	0.100	0.110	0.022	0.007 3
9	0.068	0.130	0.079	0.170	0.058	0.043 0
10	0.074	0.084	0.058	0.160	0.200	0.029 0

2. 试对长江上游望江楼1993～2000年1月份水环境监测指标(表4.13)使用最短距离聚类法作聚类分析。

要求：采用极差标准化处理原始数据，再求两两年份间的欧氏距离，最后使用最短距离聚类法作聚类分析。

表4.13 望江楼1月份水环境监测指标（1993～2000年） 单位：mg/L

年份	指标					
	DO	高锰酸盐指数	BOD_5	NH_3-N	挥发酚	镉
1993	7.8	2.9	6.2	0.47	0.009	0.000
1994	0.1	7.9	47.8	11.60	0.004	0.000
1995	2.1	9.4	31.8	3.88	0.004	0.000
1996	0.6	9.9	32.6	8.01	0.040	0.011
1997	1.6	3.9	14.9	10.50	0.014	0.000
1998	0.7	9.8	18.4	2.53	0.023	0.000
1999	0.9	10.1	41.0	2.57	0.016	0.000
2000	1.4	6.2	24.9	6.22	0.018	0.000

3. 某化工厂在附近地区挑选了有代表性的8个大气取样点，测定其中6种气体的浓度。具体数据如下表4.14，试用聚类分析分别对变量和大气污染地区进行分类评价。

要求：(1) 采用总和标准化法处理数据，然后采用夹角余弦法求6种气体（指标）间相似系数，最后选择最远距离法进行聚类分析。

(2) 采用标准差标准化法处理原始数据，采用欧氏距离测度8个样点间距离，最后用最短距离法计算类间距离，并对样本进行归类。

表4.14 各地区大气浓度值 单位：mg/L

样点	气体					
	氯	硫化氢	SO_2	碳4	环氧氯丙烷	环己烷
1	0.056	0.084	0.031	0.038	0.008 1	0.022 0
2	0.049	0.055	0.100	0.110	0.022 0	0.007 3
3	0.038	0.130	0.079	0.170	0.058 0	0.043 0
4	0.034	0.095	0.058	0.160	0.200 0	0.029 0
5	0.084	0.066	0.029	0.320	0.012 0	0.041 0
6	0.064	0.072	0.100	0.210	0.028 0	1.380 0
7	0.048	0.089	0.062	0.260	0.038 0	0.036 0
8	0.069	0.087	0.027	0.050	0.089 0	0.021 0

4. 某地区10个样地A层土壤重金属含量测定结果如下表4.15，试用最远距离聚类法对7种重金属元素进行聚类，要求聚为3类。

要求：(1) 用标准差标准化法对土壤中重金属元素的测定结果进行处理。

(2) 采用欧氏距离测度10个监测站点之间样本的距离。

(3) 选用最远距离法计算类间距离，并对样本进行归类。

表4.15　　土壤重金属测定结果　单位：mg/kg

序号	Cd	Cr	Cu	Ni	Pb	Hg	As
1	0.221	89.52	42.66	32.63	46.50	0.530	10.90
2	0.462	57.21	46.49	25.42	27.35	0.082	7.82
3	0.132	73.28	31.40	34.38	37.98	0.370	11.47
4	0.109	57.88	25.70	25.82	31.11	0.114	7.54
5	0.078	44.57	36.60	22.06	22.65	0.187	7.39
6	0.129	63.34	22.63	26.85	23.86	0.033	6.90
7	0.132	74.83	18.57	31.71	32.54	0.137	9.08
8	0.170	73.32	56.27	41.84	27.45	0.746	10.46
9	0.202	86.26	63.34	51.04	33.42	0.304	10.70
10	0.119	68.62	12.45	25.79	28.23	0.056	7.07

5. 陕西、广西、河南、江苏和云南5省的20项水安全评价指标及其取值如下表4.16(韩宇平等，2003)，试用极差的标准化方法处理数据，然后用欧氏距离法求指标间的距离，最后采用相似系数法对指标聚类。

表4.16　　水安全指标值

指　标	陕西	广西	河南	江苏	云南
C_1 人均水资源量/10^4 m^3	0.098	0.355	0.072	0.058	0.572
C_2 公顷均水资源量/10^4 m^3	0.690	3.615	0.825	0.855	3.825
C_3 地表水利用程度/%	13.157	17.619	18.407	134.752	5.565
C_4 地下水利用程度/%	28.500	3.013	41.572	10.739	0.837
C_5 工业万元产值用水量/m^3	71	192	66	81	114
C_6 农业用水综合定额/m^3	303	1 176	197	478	593

续表

指　标	陕西	广西	河南	江苏	云南
C_7 人均用水量/m^3	220	650	220	600	340
C_8 单位面积 COD 排放量/($t\cdot km^{-2}$)	1.587	4.335	4.913	6.156	0.775
C_9 工业废水处理排放达标率/%	80.880	74.000	91.520	95.890	79.120
C_{10} Ⅳ级以上水质级别占总河长比例/%	55.900	54.000	72.400	61.200	23.000
C_{11} 侵蚀模数指数	1.000	0.264	0.149	0.094	0.242
C_{12} 荒漠化指数	0.185	0.000	0.005	0.000	0.009
C_{13} 森林覆盖率指数	0.474	0.497	0.202	0.074	0.482
C_{14} 洪水受灾面积率/%	3.852	5.309	23.920	1.580	5.840
C_{15} 干旱受灾面积率/%	32.098	22.687	29.592	37.537	3.208
C_{16} 区域工农业产值密度/(10^4 元$\cdot km^{-2}$)	80.578	85.342	304.838	805.433	51.474
C_{17} 单位面积蓄水工程库容/($m^3\cdot km^{-2}$)	1.768	9.527	23.715	17.801	2.170
C_{18} 堤防保护耕地面积率/%	5.914	5.332	49.060	94.518	5.705
C_{19} 人均口粮/kg	302.108	340.499	443.118	417.666	342.304
C_{20} 粮食单产/($kg\cdot hm^{-2}$)	2 850	4 181	4 542	5 857	3 463

【参考文献】

[1] 徐建华. 计量地理学 [M]. 北京：高等教育出版社，2006.

[2] 胡永宏，贺思辉. 综合评价方法 [M]. 北京：科学出版社，2000.

[3] 韩宇平，阮本清，解建仓. 多层次多目标模糊优选模型在水安全评价中的应用 [J]. 资源科学，2003，25(4)：37-42.

第 5 章　环境模糊聚类分析

在环境数据分类中，常用的分类方法有多元统计中的系统聚类法、动态聚类法、图论聚类法等。然而现实的分类问题大多伴随模糊性，类与类之间并无清晰的界限，因此用模糊数学的方法解决这些聚类问题更为确切。模糊聚类分析已在环境科学、国民经济、社会科学、自然科学中得到广泛应用，如在环境领域中常用到的“水污染程度”的界线就是模糊的，可以采用模糊聚类分析法来解决这类问题。所谓环境模糊聚类分析就是在环境数据分类中，按照一定要求和规律对环境模糊性问题加以处理，一般先计算各样本(或变量)间的相似系数、相关系数、距离或其他表征相似程度的量来建立样本(或变量)间的模糊关系；再将模糊关系改造为模糊等价关系；然后，根据模糊等价关系，选取不同的截集，将样本分成若干类；最后完成模糊聚类分析。它的特点是：聚类的结论并不纯粹地表示对象绝对地属于某一类或绝对地不属于某一类，而是以白化的特征值表征了对象在什么程度上相对地属于某一类。其明显的用途是对所研究的环境问题的样本(或变量)进行合理的分类

本章的主要内容是：

- 模糊集理论；
- 模糊相似关系和模糊等价关系；
- 模糊聚类分析步骤；
- 传递闭包法；
- 环境应用。

5.1　模糊集理论

符合某个特定概念的全体对象，叫做该概念的外延。没有明确外延的概念，叫做模糊概念(魏世孝等，2001)。模糊现象是一种普遍存在的现象。各门学科，尤其是人文、社会学科、环境科学及其他“软科学”的数字化、定量化趋向把模糊性的数学处理问题推向中心地位。特别是计算机科学的发展，要使计算机能像人脑那样对复杂事物具有识别能力，就必须研究和处理模糊性。对于模糊性，有两种截然相反的处理方法：传统的方法是强行划清界限，人为地使每个对象都有

明确的类属，即把模糊性简化为精确性来处理；另一类方法是承认事物固有的模糊性，用元素对集合的隶属度来刻画事物从属于某类到不属于某类的逐步变化（许国志等，2001）。

1965 年，美国加利福尼亚大学控制论专家扎德（L. A. Zadeh）教授的 *Fuzzy sets* 的著名论文，宣告了“模糊数学”的正式诞生。模糊数学可称为是继经典数学和统计数学之后数学领域的又一个新发展。它用隶属函数来刻画元素对集合属于程度的连续过渡性，即元素从属于集合到不属于集合的渐变过程，中间经历了由量变到质变的连续过渡过程，也即事物具有所谓的差异中介过渡性。将经典集合的二值逻辑 {0，1} 扩展为 [0，1] 区间内的连续值逻辑，为描述和反映客观世界中各种模糊事物和现象提供了有效的手段（李荣钧，2002）。模糊数学并不是“模糊”的数学，它是采用严格的、精确的数学手段处理模糊现象的一门数学。模糊数学是传统数学的延伸、推广和补充，与传统数学一样，有着严格的数学理论基础。从认识发展的观点来看，它实际上也是对客观世界的一种精确反映，体现了人类认识能力的深化，是以模糊达到精确的手段。

5.1.1 模糊集的基本概念

模糊集合论是用隶属函数来刻画元素是否属于集合的识别过程。把被讨论的对象的全体称为论域 X，本节将简要介绍论域 X 上模糊集合的概念和表示方法。

定义 1 称 $\widetilde{A}$ 是论域 X 上的一个模糊子集（简称模糊集），如果 $\widetilde{A}$ 被一个从 X 到 [0，1] 区间的函数 $\mu_{\widetilde{A}}$ 所完全刻画：

$$\mu_{\widetilde{A}}: X \rightarrow [0, 1]$$

$$x \rightarrow \mu_{\widetilde{A}}(x), (\forall x \in X)$$

$\mu_{\widetilde{A}}$ 称为模糊子集 $\widetilde{A}$ 的隶属函数，$\mu_{\widetilde{A}}(x)$ 称为 x 隶属于 $\widetilde{A}$ 的隶属度，简记为 $\widetilde{A}(x)$。

当 $\mu_{\widetilde{A}}$ 的值域由 [0，1] 区间简化为 {0，1} 时，$\mu_{\widetilde{A}}$ 就简化为普通集合的特征函数。

5.1.2 模糊集的表示方法

在给定的论域 X 上可以有许多的模糊集，记 X 上的模糊集全体为 $F(X)$，即

$$F(X)=\{\mu_{\widetilde{A}} \mid \mu_{\widetilde{A}}: X \rightarrow [0, 1]\}$$

称 $F(X)$ 为 X 上的模糊幂集。显然普通幂集是模糊幂集的子集。

模糊集常用的表示方法有以下 4 种：

(1)向量表示法

$$\widetilde{A}=(\widetilde{A}(x_1),\ \widetilde{A}(x_2),\ \cdots,\ \widetilde{A}(x_n))$$

(2)序列表示法(序偶法)

$$\widetilde{A}=\{(x,\ \widetilde{A}(x))\mid x\in X\}=\{(x_1,\ \widetilde{A}(x_1)),\ (x_2,\ \widetilde{A}(x_2)),\ \cdots,\ (x_n,\ \widetilde{A}(x_n))\}$$

(3)分数表示法或 Zadeh 法

设论域 $X=\{x_1,\ x_2,\ \cdots,\ x_n\}$，则 X 上的模糊集可以写成：

$$\widetilde{A}=\widetilde{A}(x_1)/x_1+\widetilde{A}(x_2)/x_2+\cdots+\widetilde{A}(x_n)/x_n$$

$$\text{或 }\widetilde{A}=\widetilde{A}(x_1)/x_1\cup\widetilde{A}(x_2)/x_2\cup\cdots\cup\widetilde{A}(x_n)/x_n$$

这里的“+”或“∪”并不是求和的意思，它们只是概括集合诸元素的记号。

(4)解析法或积分表示法

当论域 X 为实数集 **R** 上的某区间时，可直接用模糊集的隶属函数的解析式来表达该模糊集。

以上的各种表示方法是当论域 X 为有限集的情况，当论域 X 为无限集的时候，X 上的模糊集可以改写成：

$$\widetilde{A}=\int_{x\in X}\mu_{\widetilde{A}}(x)/x$$

同样，这里的“$\int$”并不是求积的意思，只是概括集合诸元素的记号。

例 5.1 设论域 $X=\{x_1,\ x_2,\ x_3,\ x_4,\ x_5\}$，$x_1$，$x_2$，$x_3$，$x_4$，$x_5$ 属于“严重污染程度”分别为 0.0，0.5，0.7，0.9，1.0，则 X 上的模糊集 $\widetilde{A}=$“严重污染”的表示方法可以写成(杨晓华等，2005)：

$$\widetilde{A}=0.0/x_1+0.5/x_2+0.7/x_3+0.9/x_4+1.0/x_5$$

$$\text{或 }\widetilde{A}=\{(x_1,\ 0.0),\ (x_2,\ 0.5),\ (x_3,\ 0.7),\ (x_4,\ 0.9),\ (x_5,\ 1.0)\}$$

例 5.2 设论域 $X=\{x_1,\ x_2,\ x_3,\ x_4,\ x_5\}$，$x_1$，$x_2$，$x_3$，$x_4$，$x_5$ 分别表示环境质量评价的等级Ⅰ，Ⅱ，Ⅲ，Ⅳ，Ⅴ。某地区环境质量评价的等级属于等级Ⅰ，Ⅱ，Ⅲ，Ⅳ，Ⅴ的程度分别为 0.0，0.1，0.6，0.2，0.1，则 X 上的模糊集 $\widetilde{A}=$“某地区环境质量评价等级”的表示方法可以写成：

$$\widetilde{A}=0.0/x_1+0.1/x_2+0.6/x_3+0.2/x_4+0.1/x_5$$

$$\text{或 }\widetilde{A}=\{(x_1,\ 0.0),\ (x_2,\ 0.1),\ (x_3,\ 0.6),\ (x_4,\ 0.2),\ (x_5,\ 0.1)\}$$

例 5.3 设论域 $X=\{x_1,\ x_2,\ x_3,\ x_4,\ x_5\}$，$x_1$，$x_2$，$x_3$，$x_4$，$x_5$ 分别表示环境质量评价的 5 个指标。x_1，x_2，x_3，x_4，x_5 这 5 个指标的重要性权重分别为 0.1，0.2，0.5，0.2，0.1，则 X 上的模糊集 $\widetilde{A}=$“重要性程度”的表示方法可以写成：

$$\widetilde{A}=0.1/x_1+0.2/x_2+0.5/x_3+0.2/x_4+0.1/x_5$$

$$\text{或 }\widetilde{A}=\{(x_1,\ 0.1),\ (x_2,\ 0.2),\ (x_3,\ 0.5),\ (x_4,\ 0.2),\ (x_5,\ 0.1)\}$$

5.1.3 模糊集的运算

两模糊集之间的运算，实际上就是逐点对隶属函数作相应的运算。已出现了多种模糊运算算子(杨纶标等，2001)。目前最常用的还是 Zadeh 算子，现介绍如下。

定义 2 设$\widetilde{A}$，$\widetilde{B}\in F(U)$，若$\forall u\in U$，$\widetilde{B}(u)\leqslant\widetilde{A}(u)$

则称$\widetilde{A}$包含$\widetilde{B}$，记为$\widetilde{B}\subseteq\widetilde{A}$。若$\widetilde{B}\subseteq\widetilde{A}$，且$\widetilde{A}\subseteq\widetilde{B}$，则称$\widetilde{A}$与$\widetilde{B}$相等，记为$\widetilde{A}=\widetilde{B}$。显然，包含关系具有自反性、反对称性和传递性。

定义 3 设$\widetilde{A}$，$\widetilde{B}\in F(U)$，$\widetilde{A}$与$\widetilde{B}$的并集、交集与$\widetilde{A}$的补集(余集)分别为$\widetilde{A}\cup\widetilde{B}$，$\widetilde{A}\cap\widetilde{B}$，$\widetilde{A}^c$，它们分别由下列隶属函数完全刻画。

$$(\widetilde{A}\cup\widetilde{B})(u)=\max(\widetilde{A}(u),\ \widetilde{B}(u))=\widetilde{A}(u)\vee\widetilde{B}(u)$$

$$(\widetilde{A}\cap\widetilde{B})(u)=\min(\widetilde{A}(u),\ \widetilde{B}(u))=\widetilde{A}(u)\wedge\widetilde{B}(u)$$

$$\widetilde{A}^c(u)=1-\widetilde{A}(u)$$

例如，

$$\widetilde{A}=0.0/x_1+0.5/x_2+0.7/x_3+0.9/x_4+1.0/x_5$$

$$\widetilde{B}=0.1/x_1+0.2/x_2+0.5/x_3+0.2/x_4+0.1/x_5$$

则

$$\widetilde{A}\cup\widetilde{B}=0.1/x_1+0.5/x_2+0.7/x_3+0.9/x_4+1.0/x_5$$

$$\widetilde{A}\cap\widetilde{B}=0.0/x_1+0.2/x_2+0.5/x_3+0.2/x_4+0.1/x_5$$

$$\widetilde{A}^c=1.0/x_1+0.5/x_2+0.3/x_3+0.1/x_4+0.0/x_5$$

模糊集的并、交、补运算具有以下的性质：

(1)交换律 $\widetilde{A}\cup\widetilde{B}=\widetilde{A}\cap\widetilde{B}$；

(2)结合律 $(\widetilde{A}\cup\widetilde{B})\cup\widetilde{D}=\widetilde{A}\cup(\widetilde{B}\cup\widetilde{D})$，$(\widetilde{A}\cap\widetilde{B})\cap\widetilde{D}=\widetilde{A}\cap(\widetilde{B}\cap\widetilde{D})$；

(3)分配律 $\widetilde{A}\cup(\widetilde{B}\cap\widetilde{D})=(\widetilde{A}\cup\widetilde{B})\cap(\widetilde{A}\cup\widetilde{D})$，

$\widetilde{A}\cap(\widetilde{B}\cup\widetilde{D})=(\widetilde{A}\cap\widetilde{B})\cup(\widetilde{A}\cap\widetilde{D})$；

(4)吸收律 $\widetilde{A}\cup(\widetilde{A}\cap\widetilde{B})=\widetilde{A}$，$\widetilde{A}\cap(\widetilde{A}\cup\widetilde{B})=\widetilde{A}$；

(5)幂等律 $\widetilde{A}\cup\widetilde{A}=\widetilde{A}$，$\widetilde{A}\cap\widetilde{A}=\widetilde{A}$；

(6)对合律 $(\widetilde{A}^c)^c=\widetilde{A}$；

(7)两极律 论域U和空集$\varnothing$满足

$U\cup\widetilde{A}=U$，$U\cap\widetilde{A}=\widetilde{A}$，$\varnothing\cup\widetilde{A}=\widetilde{A}$，$\varnothing\cap\widetilde{A}=\varnothing$；

(8)对偶律 $(\widetilde{A}\cup\widetilde{B})^c=\widetilde{A}^c\cap\widetilde{B}^c$，

$(\widetilde{A}\cap\widetilde{B})^c=\widetilde{A}^c\cup\widetilde{B}^c$；

特别指出，模糊集一般不再满足互补律，即

$$\widetilde{A}\cup\widetilde{A}^c\neq U,\ \widetilde{A}\cap\widetilde{A}^c\neq\varnothing,$$

模糊集不再满足互补律，正是模糊集没有明确的边界所致。

例如，

$$\widetilde{A}=0.0/x_1+0.5/x_2+0.7/x_3+0.9/x_4+1.0/x_5$$

$$\widetilde{A}^c=1.0/x_1+0.5/x_2+0.3/x_3+0.1/x_4+0.0/x_5$$

则

$$\widetilde{A}\cup\widetilde{A}^c=1.0/x_1+0.5/x_2+0.7/x_3+0.9/x_4+1.0/x_5\neq U$$

$$\widetilde{A}\cap\widetilde{A}^c=0.0/x_1+0.5/x_2+0.3/x_3+0.1/x_4+0.0/x_5\neq\varnothing$$

5.1.4 模糊映射

定义 4 称映射 f，f：$U\to F(V)$ 为从 U 到 V 的模糊映射。即模糊映射是这样的一种对应关系，U 上的任一元素 u 与 V 上的唯一确定的模糊集对应。

例如，对于环境质量评价问题，设评价因素（指标）集 $U=\{u_1, u_2, \cdots, u_n\}$，其中 u_1，u_2，…，u_n 为被评价对象的各个因素。评价等级（评语）集 $V=\{v_1, v_2, \cdots, v_m\}$，其中 v_1，v_2，…，v_m 为各个等级（评语）。对每个单评价因素 $u_i(i=1, 2, \cdots, n)$ 进行评价，得到 V 上的模糊集 $(r_{i1}(v_1), r_{i2}(v_2), \cdots, r_{im}(v_m))$。它就是从 U 到 V 的一个模糊映射 f。

定义 5 如果 $\forall i\in N$，$j\in M$，其中，N 代表 $\boldsymbol{R}$ 矩阵的行数，M 代表 $\boldsymbol{R}$ 矩阵的列数，都有 $r_{ij}\in[0, 1]$，称矩阵 $\boldsymbol{R}=(r_{ij})_{n\times m}$ 为模糊矩阵。

例如，

$$\boldsymbol{R}=\begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.500 & 0.500 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.680 & 0.320 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}$$

就是一个模糊矩阵。

5.2 模糊关系

定义 6 给定论域 U，V 称 $U\times V$ 的一个模糊子集，$\widetilde{R}\in F(U\times V)$ 为 U 到 V 的一个模糊关系，记为 $U\xrightarrow{\widetilde{R}}V$。

设 $\tilde{R}$ 为集合 $U=\{u_1, u_2, \cdots, u_n\}$ 到 $V=(v_1, v_2, \cdots, v_m)$ 的一个模糊关系，$\forall u_i \in U$，$v_j \in V(i \in N, j \in M)$，模糊关系 $\tilde{R}$ 的隶属度 $\mu_{\tilde{R}}(u_i, v_j)$ 为 r_{ij}，则模糊关系 $\tilde{R}$ 可用如下的模糊矩阵 $\boldsymbol{R}$ 来表示：

$$\boldsymbol{R}=(r_{ij})_{n\times m}$$

其中，$r_{ij}=\mu_{\tilde{R}}(u_i, v_j)\in[0, 1]$。

对于 $u\in U$，$v\in V$，$\tilde{R}(u, v)$ 刻画了 u 对于 v 的相关程度。如果将 $\tilde{R}$ 限制为 $U\times V$ 的分明集，则此时 $\tilde{R}$ 即为普通的关系，所以模糊关系是普通关系的推广。

一般地说，由从 U 到 V 的一个模糊映射 f，可以确定一个模糊关系矩阵 $\boldsymbol{R}$。

例如，对于环境质量评价问题，对每个单评价因素 $u_i(i=1, 2, \cdots, n)$ 进行评价，得到 V 上的模糊集 $(r_{i1}, r_{i2}, \cdots, r_{im})$。它是从 U 到 V 的一个模糊映射 f，由 f 可以确定一个模糊关系矩阵 $\boldsymbol{R}$：

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{mn} \end{pmatrix}$$

例如，设 $U=\{u_1, u_2, u_3\}$ 表示父辈的三个人的集合，而 $V=\{v_1, v_2, v_3, v_4, v_5\}$ 为他们子辈的集合，$\tilde{R}\in F(U\times V)$ 表示相像关系，而且

$$\tilde{R}=\frac{0.1}{(u_1, v_1)}+\frac{0.2}{(u_1, v_2)}+\frac{0.3}{(u_2, v_1)}+\frac{0.6}{(u_2, v_2)}+\frac{0.7}{(u_2, v_3)}+\frac{0.4}{(u_3, v_1)}+\frac{0.6}{(u_3, v_2)}+\frac{0.9}{(u_3, v_4)}$$

易见 $\tilde{R}$ 是 U 到 V 的模糊关系，而 $r_{ij}=\mu_{\tilde{R}}(u_i, v_j)$ 表示 u_i 对 v_j 的相像程度，没有写出的项表示相像程度为 0，即基本上不相像。

由定义可见，模糊关系实质上是一种模糊集合，所以有关模糊集合的一切性质对其都成立。

定义 7　设 $\tilde{R}\in F(U\times V)$，$\tilde{Q}\in F(V\times W)$，则称模糊关系 $\tilde{R}\circ\tilde{Q}\in F(U\times W)$ 为 $\tilde{R}$ 与 $\tilde{Q}$ 的复合，其中 $(\tilde{R}\circ\tilde{Q})(u, w)=\bigvee_{v\in V}(\tilde{R}(u, v)\wedge\tilde{Q}(v, w))$。

若 $U=V$，而 $\tilde{R}\in F(U\times U)$，则记

$$\tilde{R}^2=\tilde{R}\circ\tilde{R},\ \tilde{R}^3=\tilde{R}^2\circ\tilde{R},\ \cdots,\ \tilde{R}^n=\tilde{R}^{n-1}\circ\tilde{R},\ \cdots$$

式中，“∨”“∧”分别为“取大”“取小”运算。

5.3　模糊等价关系

模糊等价关系的定义如下：

定义 8　设论域 U 为有限集合，U 上的一个模糊关系 R，与其对应的模糊矩阵 $\boldsymbol{R}=(r_{ij})_{n\times n}$，若满足：

(1)自反性：$r_{ii}=1$；

(2)对称性：$r_{ij}=r_{ji}$；

(3)传递性：$\boldsymbol{R}\circ\boldsymbol{R}\subseteq\boldsymbol{R}$。

则称 $\boldsymbol{R}=(r_{ij})_{n\times n}$ 是一个模糊等价矩阵，其关系是模糊等价关系。若只满足自反性和对称性则为相似关系。

例如，这里模糊相似矩阵 $\boldsymbol{R}$ 平方定义为：

$$\boldsymbol{R}\circ\boldsymbol{R}=(S_{ij})_{n\times n}$$

式中，$S_{ij}=\bigvee\limits_{k=1}^{n}(r_{ik}\wedge r_{jk})$，“∨”“∧”分别为“取大”“取小”运算。

设

$$\boldsymbol{R}=\begin{bmatrix}1.0 & 0.5 & 0.8\\ 0.5 & 1.0 & 0.5\\ 0.8 & 0.5 & 1.0\end{bmatrix}$$

显然 $\boldsymbol{R}$ 具有自反性，由

$$\boldsymbol{R}\circ\boldsymbol{R}=\begin{bmatrix}1.0 & 0.5 & 0.8\\ 0.5 & 1.0 & 0.5\\ 0.8 & 0.5 & 1.0\end{bmatrix}\circ\begin{bmatrix}1.0 & 0.5 & 0.8\\ 0.5 & 1.0 & 0.5\\ 0.8 & 0.5 & 1.0\end{bmatrix}=\begin{bmatrix}1.0 & 0.5 & 0.8\\ 0.5 & 1.0 & 0.5\\ 0.8 & 0.5 & 1.0\end{bmatrix}=\boldsymbol{R}$$

可见 $\boldsymbol{R}$ 也具有传递性，故 $\boldsymbol{R}$ 是模糊等价矩阵。

定义 9　λ 截矩阵 $\boldsymbol{R}_\lambda$：设矩阵 $\boldsymbol{R}=(r_{ij})_{n\times m}$，即：

$$\boldsymbol{R}=\begin{bmatrix}r_{11} & r_{12} & \cdots & r_{1m}\\ r_{21} & r_{22} & \cdots & r_{2m}\\ \vdots & \vdots & & \vdots\\ r_{n1} & r_{n2} & \cdots & r_{nm}\end{bmatrix}$$

记 $\boldsymbol{R}_\lambda=(r_{ij}(\lambda))_{n\times m}$。

若

$$r_{ij}(\lambda)=\begin{cases}1 & (r_{ij}\geqslant\lambda)\\ 0 & (r_{ij}<\lambda)\end{cases}$$

则称 $\boldsymbol{R}_\lambda$ 为 $\boldsymbol{R}$ 的 λ 截矩阵 $\boldsymbol{R}_\lambda$。

下面的 3 个定理是模糊聚类分析所需要的，这里只说明不证明。

定理 1　设 $\boldsymbol{R}$ 是 $U=\{u_1, u_2, \cdots, u_n\}$ 的一个自反、对称关系，即 $\boldsymbol{R}$ 是 n 阶模糊相似矩阵，则存在一个最小的自然数 $k(k\leqslant n)$，使得 $\boldsymbol{R}^k$ 为模糊等价矩阵，且对于一切大于 k 的自然数 w，恒有 $\boldsymbol{R}^w=\boldsymbol{R}^k$，$\boldsymbol{R}^k$ 称为 $\boldsymbol{R}$ 的传递包矩阵，记

为$t(\boldsymbol{R})$。

定理 2　如果模糊关系矩阵 $\boldsymbol{R}$ 是模糊等价关系，则对于任意 $\lambda \in [0, 1]$，所截的 λ 截矩阵 $\boldsymbol{R}_\lambda$ 也是等价关系。

根据这个定理，在模糊等价关系 $\boldsymbol{R}$ 确定之后，对给定的数 $\lambda \in [0, 1]$，便可得到一个相应的普通等价关系 $\boldsymbol{R}_\lambda$，可以决定一个 λ 水平分类。

定理 3　如果 $0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant 1$，则 $\boldsymbol{R}_{\lambda_2}$ 所分出的每一类必是 $\boldsymbol{R}_{\lambda_1}$ 的某一类的子类。称 $\boldsymbol{R}_{\lambda_2}$ 分类法是 $\boldsymbol{R}_{\lambda_1}$ 分类法的细化。

根据上述 3 个定理，可以进行聚类分析操作。例如，当所给矩阵关系是相似关系，由定理 1 可知，自乘若干次后，就可以获得等价关系矩阵，然后再由定理 2 和定理 3 加细分类。

5.4　模糊聚类分析步骤

模糊聚类分析步骤可以概括为：数据标准化，模糊相似矩阵的建立，聚类分析。

5.4.1 数据标准化

1. 数据标准化的作用

在实际问题中，不同的数据可能有不同的量纲。为了使不同量纲的数据也能进行比较，需要对数据进行适当的变换。根据模糊矩阵的要求将数据压缩到区间 $[0, 1]$。

2. 数据变换

设论域 $U=\{u_1, u_2, \cdots, u_n\}$ 为被分类的对象或元素，每个元素又由 m 个数据表示，对第 i 个元素有：

$$u_i=\{x_{i1}, x_{i2}, \cdots, x_{im}\} \qquad (i=1, 2, \cdots, n)$$

这时原始数据矩阵为：

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

(1) 标准差变换

$$x'_{ik}=\frac{x_{ik}-\bar{x}_k}{s_k} \qquad (i=1,2,\cdots,n;\ k=1,2,\cdots,m)$$

其中，
$$\bar{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik},\quad s_k=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$$

经过变换后，每个变量的均值为0，标准差为1，并可以消除量纲的影响，但不一定在[0，1]区间上。

（2）极差变换

$$x''_{ik}=\frac{x'_{ik}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}}{\max\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}} \qquad (k=1,2,\cdots,m)$$

经过极差变换后，消除了量纲的影响，且变换后的数据一定在[0，1]区间上。

5.4.2 模糊相似矩阵的建立

建立模糊相似矩阵，即标出衡量被分类对象间相似程度的统计量 $r_{ij}(i,j=1,2,\cdots,n)$。

设论域 $U=\{u_1,u_2,\cdots,u_n\}$，其中每个元素为一个样本，建立 U 上的相似关系 $\boldsymbol{R}$，$\boldsymbol{R}$ 表示相似矩阵 r_{ij}。每个样本为 m 维向量，$\boldsymbol{u}_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$。

计算 r_{ij} 可以有多种方法，本节仅介绍以下三种(李鸿吉，2005)。

1. 相似系数法

（1）数量积法

$$r_{ij}=\begin{cases}1 & (i=j)\\ \dfrac{1}{M}\sum\limits_{k=1}^{m}x_{ik}x_{jk} & (i\neq j)\end{cases}$$

$$M=\max_{i\neq j}\left(\sum_{k=1}^{m}x_{ik}x_{jk}\right)$$

显然 $|r_{ij}|\in[0,1]$，如果 r_{ij} 中出现负数，需要再进行变换：

$$r'_{ij}=(r_{ij}+1)/2$$

则 $r'_{ij}\in[0,1]$。

（2）夹角余弦法

$$r_{ij}=\frac{\left|\sum_{k=1}^{m}x_{ik}x_{jk}\right|}{\sqrt{\sum_{k=1}^{m}x_{ik}^{2}}\sqrt{\sum_{k=1}^{m}x_{jk}^{2}}}$$

（3）相关系数法

$$r_{ij}=\frac{\sum_{k=1}^{m}|x_{ik}-\bar{x}_i||x_{jk}-\bar{x}_j|}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)^2}}$$

2. 距离法

（1）绝对值倒数法

$$r_{ij}=\begin{cases}1 & (i=j)\\ \dfrac{M}{\sum_{k=1}^{m}|x_{ik}-x_{jk}|} & (i\neq j)\end{cases}$$

式中，M 需要适当选取，使 $0\leqslant r_{ij}\leqslant 1$。

（2）绝对值指数法

$$r_{ij}=\exp\left\{-\sum_{k=1}^{m}|x_{ik}-x_{jk}|\right\}$$

（3）直接距离法

$$r_{ij}=1-cd(\boldsymbol{u}_i,\ \boldsymbol{u}_j)$$

其中，c 为适当选取的系数，使得 $0\leqslant r_{ij}\leqslant 1$。$d$（$\boldsymbol{u}_i$，$\boldsymbol{u}_j$）为距离，经常使用的距离有以下几种。

海明距离：

$$d(\boldsymbol{u}_i,\ \boldsymbol{u}_j)=\sum_{k=1}^{m}|x_{ik}-x_{jk}|$$

欧氏距离：

$$d(\boldsymbol{u}_i,\ \boldsymbol{u}_j)=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$$

切比雪夫距离：

$$d(\boldsymbol{u}_i,\ \boldsymbol{u}_j)=\max|x_{ik}-x_{jk}|\quad(1\leqslant k\leqslant m)$$

3. 主观评分法

请专家直接对 $\boldsymbol{u}_i$ 和 $\boldsymbol{u}_j$ 的相似程度评分，也是一种有效的方法。

（1）百分制

采用百分制时，将评出的总分数除以 100，即得闭区间 [0，1] 的一个 r_{ij}。为降低主观性，可以请多个专家参与评分，再取平均，定出 r_{ij}。

（2）相似度和自信度

假定请 N 个专家组成专家组，这时有：

$$r_{ij}=\frac{\sum_{k=1}^{N} r_{ij}(k)a_{ij}(k)}{\sum_{k=1}^{N} a_{ij}(k)}$$

式中，$r_{ij}(k)$为第 k 个专家所给出的 $\boldsymbol{u}_i$ 和 $\boldsymbol{u}_j$ 的相似度，$a_{ij}(k)$是专家对自己给出的相似度的自信度。r_{ij} 和 a_{ij} 都是在 [0，1] 区间上的数值。

5.4.3 聚类分析

5.4.3.1 模糊等价矩阵聚类

1. 传递闭包法

根据所建立的模糊矩阵 $\boldsymbol{R}$，一般说来仅具有自反性和对称性，不满足传递性，只是模糊相似矩阵。只有当 $\boldsymbol{R}$ 是模糊等价矩阵时才能聚类，故需要将 $\boldsymbol{R}$ 改造成模糊等价矩阵。

由上面的定理 1 知道，可以通过求传递包将 n 阶模糊相似矩阵 $\boldsymbol{R}$ 改造成 n 阶模糊等价矩阵 $t(\boldsymbol{R})$。从模糊矩阵 $\boldsymbol{R}$ 出发，依次求平方：$\boldsymbol{R}\rightarrow\boldsymbol{R}^2\rightarrow\boldsymbol{R}^4\rightarrow\cdots$，当第一次出现 $\boldsymbol{R}^k\circ\boldsymbol{R}^k=\boldsymbol{R}^k$ 时，表明 $\boldsymbol{R}^k$ 已经具有传递性，$\boldsymbol{R}^k$ 就是所求的传递包 $t(\boldsymbol{R})$。

在 $\boldsymbol{R}$ 改造成模糊等价矩阵 $\boldsymbol{R}^k$ 之后可以在适当的限定值上进行截取，可以获得所需分类。

设论域 $U=\{u_1, u_2, u_3, u_4, u_5\}$，给定模糊关系：

$$\boldsymbol{R}=\begin{pmatrix} 1.00 & 0.50 & 0.80 & 0.40 & 0.45 \\ 0.50 & 1.00 & 0.50 & 0.40 & 0.45 \\ 0.80 & 0.50 & 1.00 & 0.40 & 0.45 \\ 0.40 & 0.40 & 0.40 & 1.00 & 0.40 \\ 0.45 & 0.45 & 0.45 & 0.40 & 1.00 \end{pmatrix}$$

其自反性和对称性是显然的，肯定是一个模糊相似矩阵。经验证可知 $\boldsymbol{R}\circ\boldsymbol{R}=\boldsymbol{R}$，故 $\boldsymbol{R}$ 又是模糊等价矩阵。根据定理 3，可以按不同水平 λ 进行分类。

（1）当 $\lambda\geqslant 1.00$ 时

此时只有对角线元素大于等于 1，故对角线元素全变成 1，其余全部为 0，成

为单位矩阵，共分为5类：$\{u_1\}$，$\{u_2\}$，$\{u_3\}$，$\{u_4\}$，$\{u_5\}$，把每一个元素分为一类，是最细的分类。

（2）当$\lambda \geqslant 0.80$时

此时小于0.80的元素都变成0，大于等于0.80的元素变成1，即有：

$$\boldsymbol{R}_\lambda = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

可以看出共分4类：$\{u_1, u_3\}$，$\{u_2\}$，$\{u_4\}$，$\{u_5\}$。

（3）当$\lambda \geqslant 0.50$时

$$\boldsymbol{R}_\lambda = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

可以看出共分3类：$\{u_1, u_2, u_3\}$，$\{u_4\}$，$\{u_5\}$。

（4）当$\lambda \geqslant 0.45$时

$$\boldsymbol{R}_\lambda = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

可以看出共分2类：$\{u_1, u_2, u_3, u_5\}$，$\{u_4\}$。

（5）当$\lambda \geqslant 0.40$时

矩阵的所有元素都变成1，只分成1类，是最粗的分类。

从上述分析可知，λ从大到小，分类从细到粗，是一个动态过程。

传递闭包法的运算量比较大，不适于手工分类，便于计算机程序设计。

2. 布尔矩阵法

设$\boldsymbol{R}$是论域$U=\{u_1, u_2, \cdots, u_n\}$上的模糊相似矩阵，若要得到$U$的元素在$\lambda$水平上的分类，使用布尔矩阵法的具体做法如下：

（1）求模糊相似矩阵$\boldsymbol{R}$的λ截矩阵$\boldsymbol{R}_\lambda$，显然$\boldsymbol{R}_\lambda$为布尔矩阵。

（2）判断$\boldsymbol{R}_\lambda$是否是等价的。如果$\boldsymbol{R}_\lambda$在任一排列下都没有下列形式的特殊子矩阵：

$$\begin{pmatrix}1&1\\1&0\end{pmatrix},\begin{pmatrix}1&1\\0&1\end{pmatrix},\begin{pmatrix}1&0\\1&1\end{pmatrix},\begin{pmatrix}0&1\\1&1\end{pmatrix}$$

则 $\boldsymbol{R}_\lambda$ 具有传递性，为等价矩阵，可以证明 $\boldsymbol{R}$ 也是等价矩阵。

(3) 如果判断 $\boldsymbol{R}_\lambda$ 是等价的，则由 $\boldsymbol{R}_\lambda$ 可得 U 在 λ 水平上的分类。

(4) 如果判断 $\boldsymbol{R}_\lambda$ 不是等价的，只要将 $\boldsymbol{R}_\lambda$ 中上述特殊形式子矩阵中的 0 一律改成 1，直到不再出现特殊形式子矩阵为止，修改后的 $\boldsymbol{R}_\lambda^*$ 为等价矩阵，可以获得 λ 水平上的分类。

5.4.3.2 直接聚类

1. 直接聚类法

在建立模糊相似矩阵后，既不求传递闭包 $t(\boldsymbol{R})$，也不用布尔矩阵法，而直接从模糊相似矩阵出发，利用相似系数进行聚类。仍用前述例子说明。

设 $U=\{u_1, u_2, u_3, u_4, u_5\}$，其模糊相似矩阵为：

$$\boldsymbol{R}=\begin{pmatrix}1.00&0.50&0.80&0.40&0.45\\0.50&1.00&0.50&0.40&0.45\\0.80&0.50&1.00&0.40&0.45\\0.40&0.40&0.40&1.00&0.40\\0.45&0.45&0.45&0.40&1.00\end{pmatrix}$$

(1) 取 $\boldsymbol{R}$ 中的最大值 $\lambda_1=1.00$(不考虑对角线元素，又由对称性 $r_{ij}=r_{ji}$，只需考虑对角线上方元素 r_{ij})，可以看出，对角线上方没有等于 1 的元素。这样，在 $\lambda_1=1.00$ 水平上的等价类为：$\{u_1\}$，$\{u_2\}$，$\{u_3\}$，$\{u_4\}$，$\{u_5\}$。

(2) 取 $\boldsymbol{R}$ 中的次大值 $\lambda_2=0.80$，由于 $r_{13}=0.80$，故相似类为 $\{u_1, u_3\}$，$\{u_2\}$，$\{u_4\}$，$\{u_5\}$。

(3) 取 $\boldsymbol{R}$ 中的第三大值 $\lambda_3=0.50$，由于 $r_{12}=r_{23}=0.50$，故相似类为 $\{u_1, u_2\}$，$\{u_2, u_3\}$，合并为等价类 $\{u_1, u_2, u_3\}$，$\{u_4\}$，$\{u_5\}$。

(4) 取 $\boldsymbol{R}$ 中的第四大值 $\lambda_4=0.45$，由于 $r_{15}=r_{25}=r_{35}=0.45$，故相似类为 $\{u_1, u_5\}$，$\{u_2, u_5\}$，$\{u_3, u_5\}$，合并为等价类 $\{u_1, u_2, u_3, u_5\}$，$\{u_4\}$。

(5) 取 $\boldsymbol{R}$ 中的最小值 $\lambda_5=0.40$，所有元素只为一类：$\{u_1, u_2, u_3, u_4, u_5\}$。

2. 最大树法

以分类元素为顶点，以相似矩阵元素 r_{ij} 为 λ，画一棵最大的树，见图 5-1。砍断低于 λ 的枝，形成一个不连贯的树枝图，各个连通的分支便构成了在 λ 水平上的分类。下面举例说明。

在讨论矩阵法分类时所获得的模糊相似矩阵为：

$$R=\begin{pmatrix} 1.00 & 0.50 & 0.80 & 0.40 & 0.45 \\ 0.50 & 1.00 & 0.50 & 0.40 & 0.45 \\ 0.80 & 0.50 & 1.00 & 0.40 & 0.45 \\ 0.40 & 0.40 & 0.40 & 1.00 & 0.40 \\ 0.45 & 0.45 & 0.45 & 0.40 & 1.00 \end{pmatrix}$$

其最大树见图5-1。

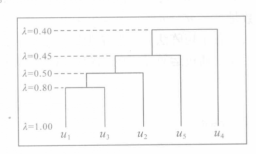


图5-1 模糊聚类图

(1) 取$\lambda=1.00$，砍去低于λ的枝，这时分5类：$\{u_1\}$，$\{u_2\}$，$\{u_3\}$，$\{u_4\}$，$\{u_5\}$；

(2) 取$\lambda=0.80$，这时分4类：$\{u_1, u_3\}$，$\{u_2\}$，$\{u_4\}$，$\{u_5\}$；

(3) 取$\lambda=0.50$，这时分3类：$\{u_1, u_2, u_3\}$，$\{u_4\}$，$\{u_5\}$；

(4) 取$\lambda=0.45$，这时分2类：$\{u_1, u_2, u_3, u_5\}$，$\{u_4\}$；

(5) 取$\lambda=0.40$，这时为1类：$\{u_1, u_2, u_3, u_4, u_5\}$。

3. 编网法

已经有了表5.1所建立的模糊相似矩阵(见“最大树法”)，例如取$\lambda=0.80$，建立λ截矩阵：

$$R_\lambda=\begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

对λ截矩阵$R_{0.80}$进行改造，方法是：对角线上换成元素名称；在对角线下方的1换成*；截矩阵中的所有的0换成空格，无论在对角线上下与否；由*向上引纵线，向右引横线(只向上和向右引线)。改造后的矩阵见图5-2。



图 5-2　由 $\lambda=0.80$ 截矩阵改造后的编网图

在图 5-2 中，* 将 u_1 和 u_3 联系起来，同样 * 将 u_2，u_4 与 u_5 也建立了联系，由此获得的分类数是 2：$\{u_1, u_3\}$，$\{u_2, u_4, u_5\}$。

我们用传递闭包法、布尔矩阵法、直接聚类法、最大树法及编网法对同一个矩阵进行模糊聚类的运算，通过运算结果的比较可以看出，这几种聚类方法得出的结论基本一致。

5.4.4　分类的 *F* 检验

从上面的一些示例可以知道，模糊聚类分析是动态的，对于不同的 $\lambda \in [0, 1]$，可以获得不同的分类。随着 λ 的变化而形成的多种分类对全面了解样本情况是有利的。但许多实际课题需要选择阈值 λ，从而给出一个较为明确的分类。用统计学的 F 检验方法可以刷掉一些不够格的类，使分类变得更为清晰。

设论域 $U=\{u_1, u_2, \cdots, u_n\}$ 是样本数为 n 的样本空间，而每个样本 u_i 有 m 个特征，记为 $\boldsymbol{u}_i=(x_{i1}, x_{i2}, \cdots, x_{im})$，由此可以得到原始数据矩阵，见表 5.1。

表 5.1　原始数据表

样本	指标					
	1	2	…	k	…	m
u_1	x_{11}	x_{12}	…	x_{1k}	…	x_{1m}
u_2	x_{21}	x_{22}	…	x_{2k}	…	x_{2m}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
u_i	x_{i1}	x_{i2}	…	x_{ik}	…	x_{im}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
u_n	x_{n1}	x_{n2}	…	x_{nk}	…	x_{nm}

总体样本的中心向量为：

$$\bar{\boldsymbol{u}} = (\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_k, \cdots, \bar{u}_m)$$

其中：

$$\bar{u}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \qquad (k=1, 2, \cdots, m)$$

设对应于 λ 值的分类数为 r。第 j 类的样本数为 n_j，其样本记为 $u_1^{(j)}$，$u_2^{(j)}$，…，$u_{n_j}^{(j)}$。第 j 类聚类中心向量为：

$$\bar{\boldsymbol{u}}^{(j)} = (\bar{u}_1^{(j)}, \bar{u}_2^{(j)}, \cdots, \bar{u}_k^{(j)}, \cdots, \bar{u}_{n_j}^{(j)})$$

式中：

$$\bar{u}_k^{(j)} = \frac{1}{n}\sum_{i=1}^{n_j} x_{ik}^{(j)} \qquad (k=1, 2, \cdots, m)$$

作 F 统计量：

$$F = \frac{\sum_{j=1}^{r} n_j \| \bar{\boldsymbol{u}}^{(j)} - \bar{\boldsymbol{u}} \|^2/(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j} \| u_i^{(j)} - \bar{u}_i^{(j)} \|^2/(n-r)}$$

其中：

$$\| \bar{\boldsymbol{u}}^{(j)} - \bar{\boldsymbol{u}} \| = \sqrt{\sum_{k=1}^{m} (\bar{u}_k^{(j)} - \bar{u}_k)^2}$$

为 $\bar{u}^{(j)}$ 与 $\bar{u}$ 的距离。

$$\| \bar{u}_i^{(j)} - \bar{u}_i^{(j)} \|$$

为第 j 类中样本 $u_i^{(j)}$ 与中心 $\bar{u}_i^{(j)}$ 的距离。

F 统计量服从自由度为 $r-1$，$n-r$ 的 F 分布。分子表征类与类之间的距离，分母表征类内样本间的距离。F 值越大，说明类与类之间的距离越大，表示类与类之间的差异大，分类明显。

表 5.2　　模糊聚类的 F 检验表

分类数	检验统计量 F 值	临界值 F_α
r	$F = \frac{\sum_{j=1}^{r} n_j \| \bar{\boldsymbol{u}}^{(j)} - \bar{\boldsymbol{u}} \|^2/(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j} \| u_i^{(j)} - \bar{u}_i^{(j)} \|^2/(n-r)}$	$F_\alpha(r-1, n-r)$
统计推断	拒绝分类数 r	$F < F_\alpha$
	接受分类数 r	$F > F_\alpha$

在一定的显著性水平下(例如 $\alpha=0.05$)，如果 $F>F_{\alpha}(r-1, n-r)$，则根据数理统计方差分析原理可以认定类与类之间的差异是显著的，说明在这样的显著性水平下分类是相对合理的。如果满足 $F>F_{\alpha}(r-1, n-r)$的分类太多，还可以提高过关的门槛，即给定更为严格的显著性水平(例如 $\alpha=0.01$)，这样可以减少分类个数。如果还是超过一个，并确实只认可一种分类，则需要从物理上考虑，由具有丰富经验的专家从物理上分析不同 λ 值的分类结果，再确定最佳分类，是一种解决这类问题的有效办法。

5.5 环境应用

下面举例说明模糊聚类分析的环境应用。

例 5.4 选取长江上游望江楼 1993～2000 年 1 月份水环境监测指标(见思考题 4 第 2 题)进行模糊聚类分析。1993 年(u_1)，1994 年(u_2)，1995 年(u_3)，1996 年(u_4)，1997 年(u_5)，1998 年(u_6)，1999 年(u_7)，2000 年(u_8)，即论域为 $U=\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$。每年度的检测指标选取 6 个主要污染物指标作为指标因子。

解 (1)将望江楼 1993～2000 年 1 月份水环境监测指标进行标准差变换：

$$x'_{ik}=\frac{x_{ik}-\bar{x}_k}{s_k} \qquad (i=1, 2, \cdots, n; k=1, 2, \cdots, m)$$

其中，

$$\bar{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}, \quad s_k=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$$

经过变换后，每个变量的均值为 0，标准差为 1，并可以消除量纲的影响，但不一定在 [0，1] 区间上。

(2) 为了将变量变换到 [0，1] 区间上，故再进行极差变换：

$$x''_{ik}=\frac{x'_{ik}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}}{\max\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}} \qquad (k=1, 2, \cdots, n)$$

式中$\max\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}$，$\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}$分别表示各同一污染因子中的最大值及最小值。经过极差变换后变量在 [0，1] 区间上，且消除了量纲的影响。极差变换后的数据，见表 5.3。

表 5.3　　标准化的数据

年份	指标					
	DO	高锰酸盐指数	BOD_5	NH_3-N	挥发酚	镉
1993	1.000 0	0.000 0	0.000 0	0.000 0	0.138 9	0.000 0
1994	0.000 0	0.694 4	1.000 0	1.000 0	0.000 0	0.000 0
1995	0.259 7	0.902 8	0.615 4	0.306 4	0.000 0	0.000 0
1996	0.064 9	0.972 2	0.634 6	0.677 4	1.000 0	1.000 0
1997	0.194 8	0.138 9	0.209 1	0.901 2	0.277 8	0.000 0
1998	0.077 9	0.958 3	0.293 3	0.185 1	0.527 8	0.000 0
1999	0.103 9	1.000 0	0.836 5	0.188 7	0.333 3	0.000 0
2000	0.168 8	0.458 3	0.449 5	0.516 6	0.388 9	0.000 0

（3）建立模糊相似矩阵

建立模糊相似矩阵又称为标定，即标出衡量分类对象间相似程度的统计量r_{ij}（i，$j=1$，2，…，n）。

采用夹角余弦法计算：

$$r_{ij}=\frac{\left|\sum_{k=1}^{m}x_{ik}x_{jk}\right|}{\sqrt{\sum_{k=1}^{m}x_{ik}^{2}}\sqrt{\sum_{k=1}^{m}x_{jk}^{2}}}$$

式中，r_{ij}为区域i与区域j的相似系数；x_{ik}，x_{jk}为两个区域的同类指标。

相似系数r_{ij}的取值范围是[0，1]，用来描述区域相似程度。如果$r_{ij}=0$，两个水质站的水污染物没有相似之处；如果$r_{ij}=1$，则两个水质站的水污染物完全相似。求出r_{ij}后，以r_{ij}为矩阵元素，即得模糊相似关系矩阵$\boldsymbol{R}$，该矩阵实际上是一个对称矩阵。

$$
\boldsymbol{R}=\begin{pmatrix}
1.000\,0 & 0.000\,0 & 0.221\,0 & 0.103\,4 & 0.232\,3 & 0.130\,2 & 0.109\,2 & 0.238\,2 \\
0.000\,0 & 1.000\,0 & 0.844\,4 & 0.646\,1 & 0.769\,7 & 0.631\,1 & 0.800\,9 & 0.879\,7 \\
0.221\,0 & 0.844\,4 & 1.000\,0 & 0.656\,8 & 0.501\,4 & 0.838\,3 & 0.947\,0 & 0.827\,4 \\
0.103\,4 & 0.646\,1 & 0.656\,8 & 1.000\,0 & 0.601\,6 & 0.790\,9 & 0.740\,8 & 0.818\,5 \\
0.232\,3 & 0.769\,7 & 0.501\,4 & 0.601\,6 & 1.000\,0 & 0.456\,9 & 0.440\,0 & 0.828\,6 \\
0.130\,2 & 0.631\,1 & 0.838\,3 & 0.790\,9 & 0.456\,9 & 1.000\,0 & 0.907\,4 & 0.830\,3 \\
0.109\,2 & 0.800\,9 & 0.947\,0 & 0.740\,8 & 0.440\,0 & 0.907\,4 & 1.000\,0 & 0.854\,4 \\
0.238\,2 & 0.879\,7 & 0.827\,4 & 0.818\,5 & 0.828\,6 & 0.830\,3 & 0.854\,4 & 1.000\,0
\end{pmatrix}
$$

（4）模糊等价矩阵聚类

经过计算，$k=4$ 时，$\boldsymbol{R}^8=\boldsymbol{R}^4$，故 $\boldsymbol{R}^8$ 就是所求的模糊等价矩阵。

$$
\boldsymbol{R}^8=\begin{pmatrix}
1.000\,0 & 0.238\,2 & 0.238\,2 & 0.238\,2 & 0.238\,2 & 0.238\,2 & 0.238\,2 & 0.238\,2 \\
0.238\,2 & 1.000\,0 & 0.854\,4 & 0.818\,5 & 0.828\,6 & 0.854\,4 & 0.854\,4 & 0.879\,7 \\
0.238\,2 & 0.854\,4 & 1.000\,0 & 0.818\,5 & 0.828\,6 & 0.907\,4 & 0.947\,0 & 0.854\,4 \\
0.238\,2 & 0.818\,5 & 0.818\,5 & 1.000\,0 & 0.818\,5 & 0.818\,5 & 0.818\,5 & 0.818\,5 \\
0.238\,2 & 0.828\,6 & 0.828\,6 & 0.818\,5 & 1.000\,0 & 0.828\,6 & 0.828\,6 & 0.828\,6 \\
0.238\,2 & 0.854\,4 & 0.907\,4 & 0.818\,5 & 0.828\,6 & 1.000\,0 & 0.907\,4 & 0.854\,4 \\
0.238\,2 & 0.854\,4 & 0.947\,0 & 0.818\,5 & 0.828\,6 & 0.907\,4 & 1.000\,0 & 0.854\,4 \\
0.238\,2 & 0.879\,7 & 0.854\,4 & 0.818\,5 & 0.828\,6 & 0.854\,4 & 0.854\,4 & 1.000\,0
\end{pmatrix}
$$

利用 $\boldsymbol{R}^8$，并根据阈值 $\lambda\in[0,1]$ 就可以进行分类。由不同的 λ 值得到的一系列分类结果，可以用动态聚类图来表示(图 5 - 3)。

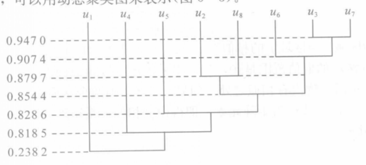


图 5 - 3　动态聚类图

由此可以得出结论：

- 当 $\lambda=0.947\,0$ 时，分 7 类：$\{u_1\}$，$\{u_2\}$，$\{u_3, u_7\}$，$\{u_4\}$，$\{u_5\}$，$\{u_6\}$，$\{u_8\}$。
- 当 $\lambda=0.907\,4$ 时，分 6 类：$\{u_1\}$，$\{u_2\}$，$\{u_3, u_6, u_7\}$，$\{u_4\}$，$\{u_5\}$，

$\{u_8\}$。

- 当 $\lambda=0.8797$ 时，分 5 类：$\{u_1\}$，$\{u_2, u_8\}$，$\{u_3, u_6, u_7\}$，$\{u_4\}$，$\{u_5\}$。
- 当 $\lambda=0.8544$ 时，分 4 类：$\{u_1\}$，$\{u_2, u_3, u_6, u_7, u_8\}$，$\{u_4\}$，$\{u_5\}$。
- 当 $\lambda=0.8286$ 时，分 3 类：$\{u_1\}$，$\{u_2, u_3, u_5, u_6, u_7, u_8\}$，$\{u_4\}$。
- 当 $\lambda=0.8185$ 时，分 2 类：$\{u_1\}$，$\{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$。
- 当 $\lambda=0.2382$ 时，分 1 类：$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$。

例 5.5　试用模糊聚类法分析评价大气环境质量。共选取了 5 个评价指标，分别是 CO、SO_2、NOx、PM_{10} 和 TSP，构造了 4 个评价等级标准，分别为：优、良、中、差。各级标准的取值范围如表 5.4 所示。实测数据，如表 5.5 所示（孙欢等，2004）。

表 5.4　　道路大气环境质量评价标准范围　单位：mg/m^3

等级	评价指标				
	CO	SO_2	NO_x	FM_{10}	TSP
优(u_1)	3.0	0.05	0.05	0.05	0.12
良(u_2)	3.5	0.10	0.08	0.10	0.20
中(u_3)	4.0	0.15	0.10	0.15	0.30
差(u_4)	6.0	0.25	0.15	0.25	0.50

表 5.5　　道路大气环境质量实测值　单位：mg/m^3

道路名称	评价指标				
	CO	SO_2	NO_x	FM_{10}	TSP
公路 $a(u_5)$	2.6	0.06	0.05	0.05	0.10
公路 $b(u_6)$	4.7	0.16	0.09	0.08	0.31
公路 $c(u_7)$	3.7	0.14	0.09	0.09	0.21
公路 $d(u_8)$	3.0	0.13	0.07	0.08	0.18

将道路大气环境质量评价标准与道路大气环境质量实测值作为整体进行模糊聚类分析，通过标准差变换和极差变换后，数据如下表 5.6 所示。

表 5.6　　标准化数据

等级/道路名称	评价指标				
	CO	SO_2	NO_x	FM_{10}	TSP
优	0.117 6	0.000 0	0.000 0	0.000 0	0.050 0
良	0.264 7	0.250 0	0.300 0	0.250 0	0.250 0
中	0.411 8	0.500 0	0.500 0	0.500 0	0.500 0
差	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
公路 *a*	0.000 0	0.050 0	0.000 0	0.000 0	0.000 0
公路 *b*	0.617 6	0.550 0	0.400 0	0.150 0	0.525 0
公路 *c*	0.323 5	0.450 0	0.400 0	0.200 0	0.275 0
公路 *d*	0.117 6	0.400 0	0.200 0	0.150 0	0.200 0

根据标准化后的数据，建立模糊相似矩阵 $\boldsymbol{R}$：

$$\boldsymbol{R}=\begin{pmatrix} 1.0000 & 0.5791 & 0.5313 & 0.5865 & 0.0000 & 0.7240 & 0.5309 & 0.3548 \\ 0.5791 & 1.0000 & 0.9944 & 0.9973 & 0.4241 & 0.9360 & 0.9707 & 0.8991 \\ 0.5313 & 0.9944 & 1.0000 & 0.9973 & 0.4623 & 0.9231 & 0.9638 & 0.9207 \\ 0.5865 & 0.9973 & 0.9973 & 1.0000 & 0.4472 & 0.9385 & 0.9657 & 0.9083 \\ 0.0000 & 0.4241 & 0.4623 & 0.4472 & 1.0000 & 0.5146 & 0.5895 & 0.7609 \\ 0.7240 & 0.9360 & 0.9231 & 0.9385 & 0.5146 & 1.0000 & 0.9581 & 0.8903 \\ 0.5309 & 0.9707 & 0.9638 & 0.9657 & 0.5895 & 0.9581 & 1.0000 & 0.9545 \\ 0.3548 & 0.8991 & 0.9207 & 0.9083 & 0.7609 & 0.8903 & 0.9545 & 1.0000 \end{pmatrix}$$

本例同样是采用在模糊等价关系基础上的聚类方法。经过计算，$k=4$ 时，$\boldsymbol{R}^8=\boldsymbol{R}^4$，故 $\boldsymbol{R}^8$ 就是所求的模糊等价矩阵。

$$\boldsymbol{R}^8=\begin{pmatrix} 1.0000 & 0.7240 & 0.7240 & 0.7240 & 0.7240 & 0.7240 & 0.7240 & 0.7240 \\ 0.7240 & 1.0000 & 0.9973 & 0.9973 & 0.7609 & 0.9581 & 0.9707 & 0.9545 \\ 0.7240 & 0.9973 & 1.0000 & 0.9973 & 0.7609 & 0.9581 & 0.9707 & 0.9545 \\ 0.7240 & 0.9973 & 0.9973 & 1.0000 & 0.7609 & 0.9581 & 0.9707 & 0.9545 \\ 0.7240 & 0.7609 & 0.7609 & 0.7609 & 1.0000 & 0.7609 & 0.7609 & 0.7609 \\ 0.7240 & 0.9581 & 0.9581 & 0.9581 & 0.7609 & 1.0000 & 0.9581 & 0.9545 \\ 0.7240 & 0.9707 & 0.9707 & 0.9707 & 0.7609 & 0.9581 & 1.0000 & 0.9545 \\ 0.7240 & 0.9545 & 0.9545 & 0.9545 & 0.7609 & 0.9545 & 0.9545 & 1.0000 \end{pmatrix}$$

利用 R^8，并根据阈值 $\lambda\in[0, 1]$就可以进行分类。λ 值根据所生成的模糊等价矩阵中的道路大气环境质量标准所在列来选取。由不同的 λ 值得到的一系列分类结果，可以用动态聚类图来表示(图 5－4)。

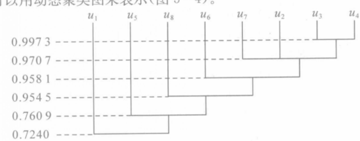


图 5－4　动态聚类图

由此可以得出结论：

- 当 $\lambda=0.9973$ 时，分 7 类：$\{u_1\}$，$\{u_2\}$，$\{u_3, u_4\}$，$\{u_5\}$，$\{u_6\}$，$\{u_7\}$，$\{u_8\}$。
- 当 $\lambda=0.9707$ 时，分 5 类：$\{u_1\}$，$\{u_2, u_3, u_4, u_7\}$，$\{u_5\}$，$\{u_6\}$，$\{u_8\}$。
- 当 $\lambda=0.9581$ 时，分 4 类：$\{u_1\}$，$\{u_2, u_3, u_4, u_6, u_7\}$，$\{u_5\}$，$\{u_8\}$。
- 当 $\lambda=0.9545$ 时，分 3 类：$\{u_1\}$，$\{u_2, u_3, u_4, u_6, u_7, u_8\}$，$\{u_5\}$。
- 当 $\lambda=0.7609$ 时，分 2 类：$\{u_1\}$，$\{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$。
- 当 $\lambda=0.7240$ 时，分 1 类：$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$。

【思考题 5】

1. 设论域 $X=\{x_1, x_2, x_3\}$，x_1，x_2，x_3 属于“严重污染程度”分别为 0.3，0.7，0.9，试给出“严重污染程度”的模糊集表示方法。

2. 设

$$R=\begin{pmatrix}1.0 & 0.5 & 0.9\\0.5 & 1.0 & 0.5\\0.9 & 0.5 & 1.0\end{pmatrix}$$

证明 R 是模糊等价矩阵。

3. 设

$$R=\begin{pmatrix}1.0000 & 0.7000 & 0.9000 & 0.9000\\0.7000 & 1.0000 & 0.7000 & 0.7000\\0.9000 & 0.7000 & 1.0000 & 0.9000\\0.9000 & 0.7000 & 0.9000 & 1.0000\end{pmatrix}$$

证明 R^2 是模糊等价矩阵。

4. 试述模糊聚类分析的详细步骤。

5. 试用传递闭包法、直接聚类法对矩阵

$$\boldsymbol{R}=\begin{pmatrix} 1.0 & 0.8 & 1.0 & 0.2 & 0.8 & 0.5 & 0.3 \\ 0.8 & 1.0 & 0.4 & 0.3 & 0.7 & 0.6 & 0.3 \\ 1.0 & 0.4 & 1.0 & 0.7 & 1.0 & 0.6 & 0.5 \\ 0.2 & 0.3 & 0.7 & 1.0 & 0.5 & 0.8 & 0.6 \\ 0.8 & 0.7 & 1.0 & 0.5 & 1.0 & 0.2 & 0.7 \\ 0.5 & 0.6 & 0.6 & 0.8 & 0.2 & 1.0 & 0.8 \\ 0.3 & 0.3 & 0.5 & 0.6 & 0.7 & 0.8 & 1.0 \end{pmatrix},$$

作聚类分析，并作 F 检验。

6. 试选一种方法对表 4.7 所示的九个农业区作聚类分析，并作 F 检验。

7. 表 5.7 为某市区 20 个采样点的地下水水质分析数据，试选一种方法对该市区采样点的地下水水质进行聚类分析，并作 F 检验。

表 5.7　　市区浅层地下水水质分析样本数据表

代号	采样地点	pH 值	Mg^{2+}	Cl^-	SO_4^{2-}	$Cl^-+SO_4^{2-}$	HCO_3^-	侵蚀性 CO_2
1	恩华药厂	7.30	37.16	123.56	138.06	261.62	259.25	0.00
2	军分区	7.50	57.39	196.37	609.99	806.36	286.06	2.64
3	兴隆大厦	7.90	40.02	73.85	486.72	560.57	317.85	0.00
4	户部商都	6.90	34.00	115.58	252.72	368.30	496.34	0.00
5	九隆总部	7.60	33.17	82.07	55.36	137.43	224.37	0.00
6	博爱大厦	7.50	104.38	503.63	19.14	522.77	245.12	0.00
7	徐州饭店	8.15	25.06	111.34	108.05	219.39	88.09	5.99
8	电力宾馆	7.65	74.32	287.78	763.01	1 050.79	442.09	0.00
9	夹河前街	7.50	57.34	141.46	176.58	318.04	358.97	0.00
10	开明市场	7.60	65.23	161.63	253.34	414.79	422.99	0.00
11	空后学院	8.40	16.43	83.42	226.38	309.80	442.58	0.00
12	灯泡厂	7.20	66.70	151.99	219.79	371.78	784.84	0.00
13	府原小区	7.70	69.25	102.53	212.40	314.93	541.80	0.00
14	天润花园	7.50	82.90	133.50	411.60	545.10	630.20	0.00
15	基督教堂	7.60	62.54	225.98	393.69	619.67	318.71	7.14
16	嬴都花园	7.55	11.95	91.03	293.46	384.49	287.79	2.43
17	交通局	7.30	35.56	41.46	64.61	106.07	443.85	0.00
18	望景花园	7.30	69.78	65.86	131.06	196.92	646.57	0.00
19	少年巷	7.60	35.44	85.05	96.31	181.36	239.32	0.00
20	体育馆	7.30	47.40	129.44	142.41	271.85	191.11	0.00

【参考文献】

[1] 魏世孝，周献中．多属性决策理论方法及其在 C^3I 系统中的应用 [M]．北京：国防工业出版社，2001.

[2] 许国志，顾基发，车宏安．系统科学 [M]．上海：上海科学技术教育出版社，2001.

[3] 李荣钧．模糊多准则决策理论与应用 [M]．北京：科学出版社，2002.

[4] 李鸿吉．模糊数学基础及实用算法 [M]．北京：科学出版社，2005.

[5] 杨纶标，高英仪．模糊数学原理及应用(第三版) [M]．广州：华南理工大学出版社，2001.

[6] 孙欢，黄川友．大气环境质量物元可拓的权值分析和实例验证 [J]．东北水力水电，2004，7(24)：42-45.

[7] 杨晓华，沈珍瑶．智能算法及其在资源环境系统建模中的应用 [M]．北京：北京师范大学出版社，2005.

第 6 章　环境判别分析

聚类分析是寻找客观分类的分析方法，而判别分析是在已知分类情况下寻找客观分析的依据。在环境科学中，我们经常遇到环境状态分类、等级评比等问题。如何根据已有分类指标或是在以往数据进行有效分类的基础上，依据研究问题的实际环境情况划分该问题所属类型，是很重要和很需要做的事情。处理这类问题的有效工具之一即是判别分析。

判别分析已成为应用性很强的一种多元统计方法。判别分析按判别的组数来分，有两组判别分析和多组判别分析；按区分不同总体所用的数学模型来分，有线性判别和非线性判别；按判别对所处理的变量方法不同又有逐步判别、序贯判别等。判别分析从不同角度提出问题，有不同的判别准则，如费歇尔(Fisher)准则和贝叶斯(Bayes)准则。本章将结合环境应用实例，介绍常用的几种判别分析方法。

本章的主要内容是：

- 距离判别分析；
- Fisher 判别分析；
- Bayes 判别分析；
- 环境应用。

6.1　距离判别分析

判别分析用统计模型的语言来描述就是：设有 k 个总体 G_1，G_2，…，G_k，希望建立一个准则，对给定的任意一个样本 x，依据这个准则就能判断它是来自哪个总体。当然，我们应当要求这种准则在某种意义下是最优先的。例如，错判概率最小或错判损失最小等。

6.1.1　两总体情况

设有两总体 G_1 和 G_2，$\boldsymbol{x}$ 是一个 p 维样本，若能定义样本 $\boldsymbol{x}$ 到总体 G_1 和 G_2

的距离 $d(\boldsymbol{x}, G_1)$ 和 $d(\boldsymbol{x}, G_2)$，则可用如下的规则进行判别：若样本 $\boldsymbol{x}$ 到总体 G_1 的距离小于到总体 G_2 的距离，则认为样本 $\boldsymbol{x}$ 属于总体 G_1；反之，则认为样本 $\boldsymbol{x}$ 属于总体 G_2；若样本 $\boldsymbol{x}$ 到总体 G_1 和 G_2 的距离相等，则让它待判。这个准则的数学模型可描述为：

$$\begin{cases} \boldsymbol{x} \in G_1 & d(\boldsymbol{x}, G_1) < d(\boldsymbol{x}, G_2) \\ \boldsymbol{x} \in G_2 & d(\boldsymbol{x}, G_1) > d(\boldsymbol{x}, G_2) \\ \text{待判} & d(\boldsymbol{x}, G_1) = d(\boldsymbol{x}, G_2) \end{cases} \tag{6.1}$$

当总体 G_1 和 G_2 为正态总体且协方差矩阵相等时，距离选用马氏距离，即：

$$d(\boldsymbol{x}, G_1) = (\boldsymbol{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \tag{6.2}$$

$$d(\boldsymbol{x}, G_2) = (\boldsymbol{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \tag{6.3}$$

这里，$\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ 分别为总体 G_1 和 G_2 的均值、协方差矩阵。$\boldsymbol{\Sigma}_1^{-1}$，$\boldsymbol{\Sigma}_2^{-1}$ 分别为 $\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ 的逆矩阵。

概括上述法则，可直观地描述为未知所属总体的样本 $\boldsymbol{x}$，离哪个总体较近，就判 $\boldsymbol{x}$ 属于哪个总体，即算出样本 $\boldsymbol{x}$ 到总体 G_1 和 G_2 的距离差，若差值为正，则样本 $\boldsymbol{x}$ 属于 G_1，否则，$\boldsymbol{x}$ 属于 G_2。

假设协方差矩阵相同，即 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$，则可证明：

$$d(\boldsymbol{x}, G_1) - d(\boldsymbol{x}, G_2) = -2[\boldsymbol{x} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2]' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

令

$$\bar{\boldsymbol{\mu}} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$$

则

$$W(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{6.4}$$

于是判别规则(6.1)可以表示为：

$$\begin{cases} \boldsymbol{x} \in G_1 & W(\boldsymbol{x}) > 0 \\ \boldsymbol{x} \in G_2 & W(\boldsymbol{x}) < 0 \\ \text{待判} & W(\boldsymbol{x}) = 0 \end{cases} \tag{6.5}$$

其中，称 $W(\boldsymbol{x})$ 为判别函数，由于它是 $\boldsymbol{x}$ 的线性函数，又称线性判别函数。线性判别的应用最为广泛，本章的大部分内容是讨论线性判别函数及其应用。

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}$ 未知时，可用样本来估计。设 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，…，$\boldsymbol{x}_{n_1}$ 是从 G_1 中取出的样本。$\boldsymbol{y}_1$，$\boldsymbol{y}_2$，…，$\boldsymbol{y}_{n_2}$ 是从 G_2 中取出的样本，则 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}$ 的估计为：

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \boldsymbol{x}_i = \bar{\boldsymbol{x}};$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \boldsymbol{y}_i = \bar{\boldsymbol{y}};$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + n_2 - 2} (A_1 + A_2)$$

其中，$A_1 = \sum_{j=1}^{n_1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$；$\boldsymbol{A}_2 = \sum_{j=1}^{n_2}(\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})'$。

如果协方差矩阵不同，即 $\boldsymbol{\Sigma}_1$ 与 $\boldsymbol{\Sigma}_2$ 不等，则判别函数 $W(\boldsymbol{x})$ 为：

$$W(\boldsymbol{x}) = d(\boldsymbol{x},\ G_1) - d(\boldsymbol{x},\ G_2)$$
$$= (\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)$$

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ 未知时，$\boldsymbol{\mu}_i$ 的估计如同方差相同时的情形，而

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{n_m - 1}A_m \qquad (m = 1,\ 2)$$

式中，
$$A_m = \sum_{i=1}^{n_m}(\boldsymbol{x}_i^{(m)} - \bar{\boldsymbol{x}}^{(m)})(\boldsymbol{x}_i^{(m)} - \bar{\boldsymbol{x}}^{(m)})'$$

当 $p=1$ 时，两总体呈图 6 - 1 的状态，这时 $W(x)$ 的符号取决于 $x > \bar{\mu}$ 还是 $x < \bar{\mu}$，从图 6 - 1 可以得到如下直观概念(何晓群，2003)：

(1) 这种判别是符合习惯的。

(2) 用这种判别方法是会发生误判的，如 x 来自总体 G_1，但却在 $\bar{\mu}$ 的右边，这时我们却判断它来自总体 G_2，误判的概率为图中阴影部分的面积(可以直观地理解成是 G_1 和 G_2 的交集)，倘若不以 $\bar{\mu}$ 为阈值点，例如以其他一点 k 来分界(图 6 - 2)，这时将总体 G_1 误判为总体 G_2 的概率是减少了，但将总体 G_2 误判为 G_1 的概率却增大了。可见，阈值点的选择是极为重要的。

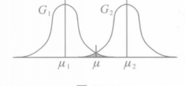


图 6 - 1

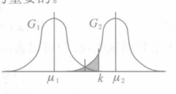


图 6 - 2

(3)如果两个总体靠得很近，则无论用何种办法，误判的概率都很大，这时勉强用判别分析意义是不大的。因此，只有当两个总体的均值有显著性差异时，作判别分析才有意义。

(4)落在 $\bar{\mu}$ 附近的样本按上述判别规则虽可进行判断，但误判可能性较大，有时划定一个待判区域，例如在此例中可定义 c 和 d，使得 $c<d$（图 6 - 3），这时判别规则改为：

图 6 - 3

$$\begin{cases} x \in G_1 & (x \leqslant c) \\ x \in G_2 & (x \geqslant d) \\ \text{待判} & (c < x < d) \end{cases}$$

综上所述，距离判别分析的步骤如下：

(1)估计总体 G_1 和 G_2 的均值、协方差矩阵；

(2)计算判别函数 $W(\boldsymbol{x})$；

(3)把待判样本代入 $W(\boldsymbol{x})$进行判断。

$$\begin{cases} \boldsymbol{x} \in G_1 & (W(\boldsymbol{x}) > 0) \\ \boldsymbol{x} \in G_2 & (W(\boldsymbol{x}) < 0) \\ \text{待判} & (W(\boldsymbol{x}) = 0) \end{cases}$$

例 6.1 根据植物的症状与受害程度来确定污染类型。假设根据叶色指数 x_1 与植株生长指数 x_2 来区分植物遭受 SO_2、HCl 等大气污染物的影响(陈玉成等，1998)。有关样本见表 6.1。试根据已知样本建立判别函数，并判定另外 3 个待判样本属于哪类。

表 6.1 两种大气污染物下的植物反应

组 别	序号	叶色指数 x_1	植株生长指数 x_2
第一组遭受 SO_2 污染	1	9.6	19.6
	2	9.3	19.9
	3	8.7	18.6
	4	8.8	18.9
	5	8.5	19.6
第二组遭受 HCl 污染	1	10.2	30.3
	2	11.3	28.7
	3	9.8	25.6
	4	7.2	27.6
	5	8.5	29.0
	6	9.6	30.0
待判样本	1	9.2	19.0
	2	8.6	19.6
	3	11.2	30.3

解 将第一组记为 G_1，第二组记为 G_2。经过计算，各类样本的指标均值为：

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{x}_i = \bar{\boldsymbol{x}} = (8.980\,0 \quad 19.320\,0)$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{y}_i = \bar{\boldsymbol{y}} = (9.4333 \quad 28.5333)$$

$$\bar{\boldsymbol{\mu}} = (9.2067 \quad 23.9267)$$

总体协方差矩阵和它的逆矩阵为：

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.2135 & 0.3317 \\ 0.3317 & 1.7979 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1} = \begin{bmatrix} 0.8678 & -0.1601 \\ -0.1601 & 0.5857 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) = (1.0817 \quad -5.3240)'$$

从而判别函数：

$$\begin{aligned} W(\boldsymbol{x}) &= (\boldsymbol{x}-\bar{\boldsymbol{\mu}})\ \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2) \\ &= 1.0817\ (x_1-9.2067)\ -5.3240\ (x_2-23.9267) \end{aligned}$$

将 3 个待判的样本数据分别代入到上面的判别函数中，可以分别求得函数值为：

$$W_1=26.2223,\ W_2=22.3789,\ W_3=-31.7753$$

$W_1>0$，$W_2>0$，$W_3<0$，根据判别函数的定义，可以判定样本 1 属于 G_1，样本 2 属于 G_1，样本 3 属于 G_2。

6.1.2 多总体情况

对应于两个总体时的情形，在分析多个总体的情况时，我们也从协方差矩阵相同和不相同两个方面来考虑。

6.1.2.1 协方差矩阵相同

设有 k 个总体 G_1，G_2，…，G_k，它们的均值分别为 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，…，$\boldsymbol{\mu}_k$，协方差矩阵均为 $\boldsymbol{\Sigma}$。类似于两总体的讨论，判别函数为：

$$W_{ij}(\boldsymbol{x}) = \left(\boldsymbol{x}-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2}\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \quad (i,\ j=1,\ 2,\ \cdots,\ k)$$

相应的判别规则为：

$$\begin{cases} \boldsymbol{x}\in G_i & (W_{ij}(\boldsymbol{x})>0,\ \forall j\neq i) \\ \text{待判} & (\text{某个 } W_{ij}(\boldsymbol{x})=0) \end{cases}$$

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，…，$\boldsymbol{\mu}_k$，$\boldsymbol{\Sigma}$ 未知时，设从 G_m 中抽取的样本为 $\boldsymbol{x}_1^{(m)}$，$\boldsymbol{x}_2^{(m)}$，…，$\boldsymbol{x}_{n_m}^{(m)}$ $(m=1,\ 2,\ \cdots,\ k)$，则它们的估计为：

$$\hat{\boldsymbol{\mu}}_m=\bar{\boldsymbol{x}}^{(m)}=\frac{1}{n_m}\sum_{i=1}^{n_m}\boldsymbol{x}_i^{(m)}\qquad(m=1,\ 2,\ \cdots,\ k)$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n-k}\sum_{m=1}^{k}A_m$$

式中：

$$n=n_1+n_2+\cdots+n_m$$

$$A_m=\sum_{i=1}^{n_m}(\boldsymbol{x}_i^{(m)}-\bar{\boldsymbol{x}}^{(m)})(\boldsymbol{x}_i^{(m)}-\bar{\boldsymbol{x}}^{(m)})'$$

6.1.2.2　协方差矩阵不相同

这时判别函数为

$$W_{ij}(\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-(\boldsymbol{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)$$

相应的判别规则为：

$$\begin{cases}\boldsymbol{x}\in G_i & (W_{ij}(\boldsymbol{x})>0,\ \forall j\neq i)\\ \text{待判} & (\text{某个 } W_{ij}(\boldsymbol{x})=0)\end{cases}$$

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，…，$\boldsymbol{\mu}_k$，$\boldsymbol{\Sigma}_1$，…，$\boldsymbol{\Sigma}_k$ 未知时，$\boldsymbol{\mu}_i$ 的估计如同方差相同时的情形，而

$$\hat{\boldsymbol{\Sigma}}_m=\frac{1}{n_m-1}A_m\qquad(m=1,\ 2,\ \cdots,\ k)$$

式中：

$$A_m=\sum_{i=1}^{n_m}(\boldsymbol{x}_i^{(m)}-\bar{\boldsymbol{x}}^{(m)})(\boldsymbol{x}_i^{(m)}-\bar{\boldsymbol{x}}^{(m)})'$$

6.2　Fisher 判别

同方差分析思路相类似，Fisher 判别准则是寻找一种判别函数，使类间均值差的平方和与类内差异平方和之比为极大值。Fisher 判别也可用于多类判别，但在国内外，多类判别更多的是采用贝叶斯(Bayes)判别准则。

假设某个环境问题是由 p 个因子(变量)组成，现在欲构造一个似然函数：

$$L(x_1,\ x_2,\ \cdots,\ x_p)=y=c_1x_1+c_2x_2+\cdots+c_px_p$$

其中，x_i 代表第 i 个变量，c_i 是变量的系数。假设有两类样本，属于第一类的有 n_1 个，属于第二类的有 n_2 个，它们的数据矩阵如下：

$$\boldsymbol{X}^{(1)}=\begin{pmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1n_1}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2n_1}^{(1)} \\ \vdots & \vdots & & \vdots \\ x_{p1}^{(1)} & x_{p2}^{(1)} & \cdots & x_{pn_1}^{(1)} \end{pmatrix}$$

$$\boldsymbol{X}^{(2)}=\begin{pmatrix} x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1n_2}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2n_2}^{(2)} \\ \vdots & \vdots & & \vdots \\ x_{p1}^{(2)} & x_{p2}^{(2)} & \cdots & x_{pn_2}^{(2)} \end{pmatrix}$$

$\boldsymbol{X}$ 右上角的数码表明它们分别属于第一、第二类样本。我们规定某一个定值 y_0，当把某样本的 p 项指标代入上述函数 y 中，如果有 $y>y_0$，我们判定它属于第一类；反之，则属于第二类。现在假设我们已经找到符合上述判别要求的判别函数，并把已知分属于第一、二类样本的数据代入式子中，则有：

$$y_j^{(1)}=c_1x_{1j}^{(1)}+c_2x_{2j}^{(1)}+\cdots+c_px_{pj}^{(1)} \qquad (j=1,\ 2,\ \cdots,\ n_1)$$
$$y_j^{(2)}=c_1x_{1j}^{(2)}+c_2x_{2j}^{(2)}+\cdots+c_px_{pj}^{(2)} \qquad (j=1,\ 2,\ \cdots,\ n_2)$$

令：

$$\overline{y^{(1)}}=\frac{1}{n_1}\sum_{j=1}^{n_1}y_j^{(1)},\quad \overline{y^{(2)}}=\frac{1}{n_2}\sum_{j=1}^{n_2}y_j^{(2)}$$

对于分类来说,显然要求：

(1) $Q=(\overline{y^{(1)}}-\overline{y^{(2)}})^2$ 越大越好，即类间均值差越大越好；

(2)希望类内差异越小越好，即 $\sum_{j=1}^{n_1}(y_j^{(1)}-\overline{y^{(1)}})^2$ 和 $\sum_{j=1}^{n_2}(y_j^{(2)}-\overline{y^{(2)}})^2$ 越小越好。记：

$$F=\sum_{j=1}^{n_1}(y_j^{(1)}-\overline{y^{(1)}})^2+\sum_{j=1}^{n_2}(y_j^{(2)}-\overline{y^{(2)}})^2$$

根据上述两点准则，我们应该使

$$I=\frac{Q}{F}$$

越大越好。

对上式两边分取对数有：

$$\ln I=\ln Q-\ln F$$

显然当 I 取得极大值时，$\ln I$ 也取得极大值。欲求系数 $c_i(i=1,\ 2,\ \cdots,\ p)$满足极值条件，应有：

$$\frac{\partial \ln I}{\partial c_i}=\frac{\partial \ln Q}{\partial c_i}-\frac{\partial \ln F}{\partial c_i}=0$$

即：

$$\frac{1}{Q}\frac{\partial Q}{\partial c_i}-\frac{1}{F}\frac{\partial F}{\partial c_i}=0$$

由此有：

$$\frac{\partial Q}{\partial c_i}=\frac{Q}{F}\frac{\partial F}{\partial c_i}=I\frac{\partial F}{\partial c_i}$$

由于：

$$\begin{aligned}\overline{y^{(1)}} &= \frac{1}{n_1}\sum_{j=1}^{n_1} y_j^{(1)} = \frac{1}{n_1}\sum_{j=1}^{n_1}(c_1x_{1j}^{(1)}+c_2x_{2j}^{(1)}+\cdots+c_px_{pj}^{(1)}) \\ &= c_1\,\overline{x_1^{(1)}}+c_2\,\overline{x_2^{(1)}}+\cdots+c_p\,\overline{x_p^{(1)}} \\ \overline{y^{(2)}} &= c_1\,\overline{x_1^{(2)}}+c_2\,\overline{x_2^{(2)}}+\cdots+c_p\,\overline{x_p^{(2)}}\end{aligned}$$

从而：

$$\begin{aligned}F &= \sum_{j=1}^{n_1}(y_j^{(1)}-\overline{y^{(1)}})^2+\sum_{j=1}^{n_2}(y_j^{(2)}-\overline{y^{(2)}})^2 \\ &= \sum_{j=1}^{n_1}\left[\sum_{k=1}^{p}c_k(x_{kj}^{(1)}-\overline{x_k^{(1)}})\right]^2+\sum_{j=1}^{n_2}\left[\sum_{k=1}^{p}c_k(x_{kj}^{(2)}-\overline{x_k^{(2)}})\right]^2 \\ &= \sum_{j=1}^{n_1}\left[\sum_{k=1}^{p}c_k(x_{kj}^{(1)}-\overline{x_k^{(1)}})\sum_{k=1}^{p}c_k(x_{kj}^{(1)}-\overline{x_k^{(1)}})\right]+ \\ &\quad \sum_{j=1}^{n_2}\left[\sum_{k=1}^{p}c_k(x_{kj}^{(2)}-\overline{x_k^{(2)}})\sum_{k=1}^{p}c_k(x_{kj}^{(2)}-\overline{x_k^{(2)}})\right] \\ &= \sum_{j=1}^{n_1}\left[\sum_{k=1}^{p}c_k(x_{kj}^{(1)}-\overline{x_k^{(1)}})\sum_{e=1}^{p}c_e(x_{ej}^{(1)}-\overline{x_e^{(1)}})\right]+ \\ &\quad \sum_{j=1}^{n_2}\left[\sum_{k=1}^{p}c_k(x_{kj}^{(2)}-\overline{x_k^{(2)}})\sum_{e=1}^{p}c_e(x_{ej}^{(2)}-\overline{x_e^{(2)}})\right] \\ &= \sum_{j=1}^{n_1}\sum_{k=1}^{p}\sum_{e=1}^{p}c_kc_e(x_{kj}^{(1)}-\overline{x_k^{(1)}})(x_{ej}^{(1)}-\overline{x_e^{(1)}})+ \\ &\quad \sum_{j=1}^{n_2}\sum_{k=1}^{p}\sum_{e=1}^{p}c_kc_e(x_{kj}^{(2)}-\overline{x_k^{(2)}})(x_{ej}^{(2)}-\overline{x_e^{(2)}}) \\ &= \sum_{k=1}^{p}\sum_{e=1}^{p}c_kc_el_{ke}\end{aligned}$$

这里：

$$l_{ke}=\sum_{j=1}^{n_1}(x_{kj}^{(1)}-\overline{x_k^{(1)}})(x_{ej}^{(1)}-\overline{x_e^{(1)}})+\sum_{j=1}^{n_2}(x_{kj}^{(2)}-\overline{x_k^{(2)}})(x_{ej}^{(2)}-\overline{x_e^{(2)}})$$

显然：

$$l_{ke}=l_{ek}$$

并且有：

$$\sum_{j=1}^{n_1}(x_{kj}^{(1)}-\overline{x_k^{(1)}})(x_{ej}^{(1)}-\overline{x_e^{(1)}})=\sum_{j=1}^{n_1}x_{kj}^{(1)}x_{ej}^{(1)}-\frac{1}{n_1}\sum_{j=1}^{n_1}x_{kj}^{(1)}\sum_{j=1}^{n_2}x_{ej}^{(1)}$$

当 $k=e$ 时，上式 $=\sum_{j=1}^{n_1}(x_{kj}^{(1)})^2-\frac{1}{n_1}\left[\sum_{j=1}^{n_1}x_{kj}^{(1)}\right]^2$

对 $x^{(2)}$ 也有同样结果。

所以有：

$$\frac{\partial F}{\partial c_k}=2(c_1l_{k1}+c_2l_{k2}+\cdots+c_pl_{kp})\qquad(k=1,\ 2,\ \cdots,\ p)$$

$$Q=(\overline{y^{(1)}}-\overline{y^{(2)}})^2=\left[c_1(\overline{x_1^{(1)}}-\overline{x_1^{(2)}})+c_2(\overline{x_2^{(1)}}-\overline{x_2^{(2)}})+\cdots+c_p(\overline{x_p^{(1)}}-\overline{x_p^{(2)}})\right]^2$$

$$=\left[\sum_{e=1}^{p}c_e(\overline{x_e^{(1)}}-\overline{x_e^{(2)}})\right]^2=\left(\sum_{e=1}^{p}c_et_e\right)^2$$

这里：

$$t_e=\overline{x_e^{(1)}}-\overline{x_e^{(2)}}\qquad(e=1,\ 2,\ \cdots,\ p)$$

由此有：

$$\frac{\partial Q}{\partial c_k}=2\left(\sum_{e=1}^{p}c_et_e\right)t_k\qquad(k=1,\ 2,\ \cdots,\ p)$$

$$I\frac{\partial F}{\partial c_k}=\frac{\partial Q}{\partial c_k}$$

$$I(c_1l_{k1}+c_2l_{k2}+\cdots+c_pl_{kp})=\left(\sum_{e=1}^{p}c_et_e\right)t_k\qquad(k=1,\ 2,\ \cdots,\ p)$$

令：

$$\beta=\frac{\sum_{e=1}^{p}c_et_e}{I}$$

则有：

$$c_1l_{k1}+c_2l_{k2}+\cdots+c_pl_{kp}=\beta t_k$$

因而有：

$$
\begin{cases}
c_1 l_{11} + c_2 l_{12} + \cdots + c_p l_{1p} = \beta t_1 \\
c_1 l_{21} + c_2 l_{22} + \cdots + c_p l_{2p} = \beta t_2 \\
\cdots \\
c_1 l_{p1} + c_2 l_{p2} + \cdots + c_p l_{pp} = \beta t_p
\end{cases}
$$

上面方程组的解是如下方程组解的 β 倍：

$$\sum_{e=1}^{p} c_e l_{ke} = t_k \quad (k=1,\ 2,\ \cdots,\ p)$$

由 $I=\frac{Q}{F}$，Q，F 的定义知上述解(βc_1，βc_2，…，βc_p)，代入方程时正好可以把 β 提取并相互抵消。所以方程组 $\sum_{e=1}^{p} c_e l_{ke} = t_k (k=1,\ 2,\ \cdots,\ p)$ 的解 $c_e(e=1,\ 2,\ \cdots,\ p)$ 可使 I 达到最大值。这样，作为线性函数：

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$$

就是我们所要求的，被称为判别函数。

实际上，可以证明，$\boldsymbol{c}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$。实际进行判别时，常令 $\beta=1$ 或 $\beta=n_1+n_2-2$。因而 y_0 的选择有：

$$y_0 = \frac{n_1 \overline{y^{(1)}} + n_2 \overline{y^{(2)}}}{n_1 + n_2}$$

它是 $\overline{y^{(1)}}$，$\overline{y^{(2)}}$ 的加权平均。y_0 也可以用如下方法表示：

$$y_0 = \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)$$

不失一般性，假设：

$$\overline{y^{(1)}} > y_0 > \overline{y^{(2)}}$$

这样当某一待判样本的数据代入判别函数时，如果其值 $y > y_0$，则属于第一类；否则属于第二类。

Fisher 判别分析的步骤如下：

(1) 估计总体 G_1,G_2 的均值和协方差矩阵 $\hat{\boldsymbol{\mu}}_1$，$\hat{\boldsymbol{\mu}}_2$，$\hat{\boldsymbol{\Sigma}}$；

(2) 计算 $y_0 = \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)$；

(3) 计算 $\boldsymbol{c} = \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$；

(4) 求判别函数 $y = \boldsymbol{x}'\boldsymbol{c}$；

(5)计算判别函数分类值 $\overline{y^{(1)}} = \frac{1}{n_1}\sum_{j=1}^{n_1} y_j^{(1)}$，$\overline{y^{(2)}} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j^{(2)}$；

(6) 待判样本的判别。

先计算待判样本的判别函数值，然后对照其与 y_0，$\overline{y^{(1)}}$，$\overline{y^{(2)}}$ 的大小，最后确定待判样本的归属。

例 6.2 试对第 1 节例 6.1 中的问题用 Fisher 判别法判断。

解 根据距离判别中的分析数据，可以得到：

$$y_0=\frac{1}{2}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)=-117.4263$$

$$y=\boldsymbol{c}'\boldsymbol{x}=(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}=1.0817x_1-5.3240x_2$$

$$\overline{y^{(1)}}=\frac{1}{n_1}\sum_{j=1}^{n_1}y_j^{(1)}=-93.1457$$

$$\overline{y^{(2)}}=\frac{1}{n_2}\sum_{j=1}^{n_2}y_j^{(2)}=-141.7070$$

即 $\overline{y^{(1)}}>y_0>\overline{y^{(2)}}$

将样本 1，2，3 的数据分别代入到判别函数中，得到：

$$y_1=-91.2040,\ y_2=-95.0474,\ y_3=-149.2017$$

根据 Fisher 判别准则，可以判定样本 1 属于 G_1，样本 2 属于 G_1，样本 3 属于 G_2。这个结果和距离判别的结果是一致的。

6.3 Bayes 判别

简单地说，判别分析就是根据掌握的历史上每个类别的若干样本的数据信息，总结出客观事物分类的规律性，建立判别函数和判别准则，然后，当遇到新的样本点时，只需根据总结出来的判别函数和判别准则，就能够判别该样本点的所属类别。

除了上述 Fisher 判别准则外，解决判别问题的另一条途径是 Bayes 判别分析。

设 G_1 和 G_2 是两个总体。在一次发生的事件中，人们先验的给出该事件属于 G_1 总体的概率为 $q_1=P(G_1)$，属于 G_2 总体的概率为 $q_2=P(G_2)$，这些概率称为先验概率。如果事件发生在 G_1 总体中或是 G_2 总体中是互不相容的，则应该有：

$$q_1+q_2=1$$

同时，我们把在已知总体 $G_i(i=1,2)$条件下观测到事件 x 的概率称为条件概率，记为 $P(x\mid G_i)$，把已知事件发生并知道它来自总体 $G_i(i=1,2)$时的概率称为后验概率，记为 $P(G_i\mid x)$。这样 Bayes 公式为：

$$P(G_i \mid x)=\frac{P(G_i)P(x \mid G_i)}{\sum_{i=1}^{2}P(x \mid G_i)P(G_i)}=\frac{q_iP(x \mid G_i)}{q_1P(x \mid G_1)+q_2P(x \mid G_2)} \quad (i=1,\ 2) \tag{6.6}$$

很自然，人们能设想，对任意一个事件 x，若 $P(G_1 \mid x) \geqslant P(G_2 \mid x)$，则事件 x 属于总体 G_1；反之，若 $P(G_1 \mid x)<P(G_2 \mid x)$，事件属于总体 G_2。当总体为多元正态分布时，上式中 $P(x \mid G_i)$ 可用概率密度 $p_i(x)$ 代替，即有：

$$P(G_i \mid x)=\frac{q_ip_i(x)}{q_1p_1(x)+q_2p_2(x)} \quad (i=1,\ 2) \tag{6.7}$$

把上式代入 $P(G_1 \mid x) \geqslant P(G_2 \mid x)$，$P(G_1 \mid x)<P(G_2 \mid x)$ 就有：

若 $\frac{q_1p_1(x)}{q_2p_2(x)} \geqslant 1$ 事件 x 划归 G_1 总体；

若 $\frac{q_1p_1(x)}{q_2p_2(x)} < 1$ 事件 x 划归 G_2 总体。

可以证明，按照 Bayes 判别准则可使“错分的平均概率”

$$q_1P(2 \mid 1)+q_2P(1 \mid 2) \tag{6.8}$$

达到极小值。这里 $P(2 \mid 1)$ 表示某一事件属于总体 1 却错划为总体 2 的概率；$P(1 \mid 2)$ 表示某一事件属于总体 2 却错划为总体 1 的概率。

进一步地，如果记总体是 G_1 发生的事件却划分在总体 G_2 的损失为 $L(2 \mid 1)$，而总体是 G_2 发生的事件却记在总体 G_1 上的损失为 $L(1 \mid 2)$。那么，“错分平均损失”定义为：

$$q_1L(2 \mid 1)P(2 \mid 1)+q_2L(1 \mid 2)P(1 \mid 2) \tag{6.9}$$

下面，我们把上述 Bayes 二类判别准则推广到多类判别。

考虑 m 个总体 G_1，G_2，…，G_m，每个总体均是 p 维，且具有概率分布密度 $p_1(\boldsymbol{x})$，$p_2(\boldsymbol{x})$，…，$p_m(\boldsymbol{x})$。任意取一样本 $\boldsymbol{x}$，它属于 G_i 的先验概率为 $q_i(i=1,\ 2,\ \cdots,\ m)$。根据这个，我们来考虑 $\boldsymbol{x}$ 的归属问题。

样本 $\boldsymbol{x}$ 是 p 维空间中的一点，如果该空间被对应于 G_1，G_2，…，G_m 的总体分割成 m 个空间 D_1，D_2，…，D_m。当样本落入某 D_i 空间，样本就判属于总体 G_i。若 $\boldsymbol{x}$ 属于 G_i 却判定属于 G_j，将它带来的损失定义为：

$$P(j \mid i)=\int_{D_j} p_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (i \neq j,\ j=1,\ 2,\ \cdots,\ m)$$

因为上述划分 D_1，D_2，…，D_m 进行判别而造成的平均损失为：

$$g(D_1,\ D_2,\ \cdots,\ D_m)=\sum_{i=1}^{m}q_i\sum_{j=1}^{m}L(j \mid i)P(j \mid i) \tag{6.10}$$

可以证明，当某种划分 D_1，D_2，…，D_m 满足：

$$D_l=\{x: h_l(x)<h_j(x)\quad (j\neq l;\ j=1,2,\cdots,m;\ l=1,2,\cdots,m)\}$$

时，能使平均损失 $g(D_1, D_2, \cdots, D_m)$取得最小值。其中：

$$h_j(\boldsymbol{x})=\sum_{\substack{i=1\\i\neq j}}^{m} q_i L(j\mid i) p_i(\boldsymbol{x}) \tag{6.11}$$

这就是说，对任意一个样本，计算出 $h_i(\boldsymbol{x})(i=1, 2, \cdots, m)$。若在所有的 $h_i(\boldsymbol{x})$中，$h_l(\boldsymbol{x})$最小，则判断 $\boldsymbol{x}$ 属于 D_l，因而 $\boldsymbol{x}$ 划归于 G_l 类。在实际中，式(6.11)并不容易计算出，所以往往假设 $L(j\mid i)=1$。这样：

$$h_l(\boldsymbol{x})=\sum_{\substack{i=1\\i\neq l}}^{m} q_i p_i(\boldsymbol{x})=\sum_{i=1}^{m} q_i p_i(\boldsymbol{x})-q_l p_l(\boldsymbol{x})=1-q_l p_l(\boldsymbol{x}) \tag{6.12}$$

$h_l(\boldsymbol{x})$极小，则 $q_l p_l(\boldsymbol{x})$极大。由此知，若 $\boldsymbol{x}$ 划归 D_l 的概率最大，则 $\boldsymbol{x}$ 就应该判属于 G_l 类。因而，就产生了寻找判别 $q_l p_l(\boldsymbol{x})$最大值的函数。

可以证明，当总体是 p 维正态分布时，第 i 个总体分布密度为：

$$p_i(\boldsymbol{x})=\frac{|\boldsymbol{\Sigma}^{-1}|^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \tag{6.13}$$

这里，$\boldsymbol{\Sigma}$ 是总体协方差矩阵，$\boldsymbol{\Sigma}^{-1}$ 是它的逆矩阵，μ_i 是第 i 个总体均值向量，$\boldsymbol{x}$ 则为样本向量。当先验概率 q_i 和 μ_i 均已知时，可以建立判别函数：

$$y_i(\boldsymbol{x})=c_{0i}+c_{1i}x_1+c_{2i}x_2+\cdots+c_{pi}x_p+\ln q_i \quad (i=1, 2, \cdots, m) \tag{6.14}$$

这里，c_{ji} 称为判别系数，c_{0i} 为常数项。这样可以得到如下样本归属判断准则：

设 $G_i\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$，把样本观测值代入式(6.14)得到值 $y_i(\boldsymbol{x})$，它的先验概率为 q_i，错分损失为 $L(j\mid i)(j\neq i)$，相应的 Bayes 解为：

$$D_i=(\boldsymbol{x}\mid y_i(\boldsymbol{x})=\max_{i,j=1,2,\cdots,m} y_j(\boldsymbol{x}))$$

因此，当样本观测值代入 m 个线性回归判别函数 $y_i(i=1, 2, \cdots, m)$，计算并进行比较，得最大 $y_i(\boldsymbol{x})$，样本就归属于 G_i 类。

在实际问题中，$\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Sigma}$ 往往是未知的，多以样本均值 $\bar{x}_i$ 和样本协方差矩阵 $\boldsymbol{L}$ 作为 $\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Sigma}$ 的估计值，并且在计算中常常令 $q_1=q_2=\cdots=q_m=1$，所以判别函数有较为简单的形式：

$$y_i(\boldsymbol{x})=c_{0i}+c_{1i}x_1+c_{2i}x_2+\cdots+c_{pi}x_p \quad (i=1, 2, \cdots, m)$$

具体在计算系数时，可以证明：

$$y_i=\ln q_i+\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \quad (i=1, 2, \cdots, m) \tag{6.15}$$

其中，$\boldsymbol{\Sigma}=\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})'$，$n$ 为所有的各类样本个数，n_i

为第 i 类的样本数。

Bayes 判别分析的步骤如下：

(1)估计总体的均值和协方差阵；

(2)求 m 个判别函数 $y_i=\ln q_i+\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \quad (i=1, 2, \cdots, m)$；

(3)待判样本的判别。

先计算待判样本的 m 个判别函数值，然后进行比较，找到最大的 y_i，这样本就归属于 G_i 类。

6.4　环境应用

例 6.3　贝叶斯判别分析理论在安全评价中的应用

南方某矿业集团，对其下属企业的 3 个矿井中的环境条件因素进行评价，根据南方煤矿的特点和以往的经验，采取的评价因素为：X_1：巷道合格率/%；X_2：粉尘浓度/$mg \cdot m^{-3}$；X_3：环境温度/℃；X_4：风速/$m \cdot s^{-1}$；X_5：巷道最小行人宽度/m；X_6：巷道最小行人高度/m，共 6 项综合指标进行评判，其原始数据(采用评分法)和指标体系，安全评价等级见表 6.2 和表 6.3(雷兢等，2004)。试用上述贝叶斯判别原理，建立安全评价等级函数。

表 6.2　　安全评价等级表

指标	安全评价等级				
	安全(5)	较安全(4)	一般安全(3)	较不安全(2)	不安全(1)
巷道合格率/%	>95	90～95	85～90	80～85	< 80
粉尘浓度/($mg \cdot m^{-3}$)	< 4	4～6	6～8	8～10	>10
环境温度/℃	18～22	22～24	24～26	26～28	>28
风速/($m \cdot s^{-1}$)	2.5～3.5	2.0～2.5	1.5～2.0	1.5～1.0	<1.0
巷道最小行人宽度/m	>1.2	1.1～1.2	1.0～1.1	0.8～1.0	<0.8
巷道最小行人高度/m	>1.8	1.6～1.8	1.4～1.6	1.2～1.4	<1.2

表 6.3　原始数据

	矿井编号	X_1	X_2	X_3	X_4	X_5	X_6	期望值
原始样本	1	97.38	2.12	21.5	2.87	1.40	1.83	5
	2	98.10	3.65	19.5	3.35	1.31	2.24	5
	3	96.45	3.14	18.0	3.50	1.20	1.94	5
	4	95.30	3.87	22.0	2.56	1.25	2.50	5
	5	94.87	4.03	23.1	2.01	1.17	1.79	4
	6	93.15	5.35	22.7	2.32	1.19	1.72	4
	7	91.57	4.89	22.2	2.21	1.13	1.68	4
	8	90.78	5.87	23.8	2.48	1.10	1.60	4
	9	87.69	6.17	25.9	1.98	1.05	1.47	3
	10	89.34	7.32	24.3	1.55	1.07	1.52	3
	11	85.10	6.87	25.2	1.63	1.09	1.59	3
	12	86.54	7.91	24.0	1.75	1.00	1.41	3
	13	84.68	8.07	26.1	1.50	0.86	1.39	2
	14	84.10	9.13	27.9	1.48	0.91	1.28	2
	15	82.34	8.63	27.4	1.35	0.97	1.35	2
	16	80.25	9.87	26.5	1.14	0.81	1.20	2
	17	75.68	10.05	30.7	0.56	0.79	1.89	1
	18	78.98	12.30	28.9	0.87	0.45	0.78	1
	19	73.56	11.28	29.6	0.96	0.58	0.98	1
	20	70.14	10.87	28.5	0.48	0.74	1.17	1

解　根据表中的数据，得到：

$$\hat{\boldsymbol{\mu}}_1=(96.8075\quad 3.1950\quad 20.2500\quad 3.0700\quad 1.2900\quad 2.1275)$$

$$\hat{\boldsymbol{\mu}}_2=(92.5925\quad 5.0350\quad 22.9500\quad 2.2550\quad 1.1475\quad 1.6975)$$

$$\hat{\boldsymbol{\mu}}_3=(87.1675\quad 7.0675\quad 24.8500\quad 1.7275\quad 1.0525\quad 1.4975)$$

$$\hat{\boldsymbol{\mu}}_4=(82.8425\quad 8.9250\quad 26.9750\quad 1.3675\quad 0.8875\quad 1.3050)$$

$$\hat{\boldsymbol{\mu}}_5=(74.5900\quad 11.1250\quad 29.4250\quad 0.7175\quad 0.6400\quad 1.2050)$$

$$\hat{\boldsymbol{\Sigma}}=\begin{pmatrix} 5.1429 & -0.2132 & 0.0066 & 0.1467 & -0.0376 & -0.0374 \\ -0.2132 & 0.6430 & -0.2012 & 0.0233 & -0.0474 & -0.0697 \\ 0.0066 & -0.2012 & 1.2457 & -0.1179 & 0.0381 & 0.1122 \\ 0.1467 & 0.0233 & -0.1179 & 0.0687 & -0.0088 & -0.0288 \\ -0.0376 & -0.0474 & 0.0381 & -0.0088 & 0.0079 & 0.0127 \\ -0.0374 & -0.0697 & 0.1122 & -0.0288 & 0.0127 & 0.0689 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}=\begin{pmatrix} 0.2379 & 0.2675 & -0.0700 & -0.4287 & 2.9515 & -0.2091 \\ 0.2675 & 3.2278 & -0.0498 & 1.0348 & 22.5338 & -0.2288 \\ -0.0700 & -0.0498 & 1.0824 & 1.3111 & -3.2848 & -0.6976 \\ -0.4287 & 1.0348 & 1.3111 & 21.1819 & 13.3569 & 5.0551 \\ 2.9515 & 22.5338 & -3.2848 & 13.3569 & 357.7665 & -30.5919 \\ -0.2091 & -0.2288 & -0.6976 & 5.0551 & -30.5919 & 23.0396 \end{pmatrix}$$

$$\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}=(24.5171\quad 66.9584\quad 13.2904\quad 81.3633\quad 728.6500\quad -10.0293)'$$

$$\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}=(23.8363\quad 67.6796\quad 16.1156\quad 67.2740\quad 700.0895\quad -21.1198)'$$

$$\hat{\boldsymbol{\mu}}_3'\hat{\boldsymbol{\Sigma}}^{-1}=(22.9439\quad 70.0536\quad 18.2104\quad 60.7408\quad 688.7211\quad -26.1440)'$$

$$\hat{\boldsymbol{\mu}}_4'\hat{\boldsymbol{\Sigma}}^{-1}=(21.9707\quad 70.7400\quad 20.9248\quad 56.5007\quad 652.8811\quad -28.3541)'$$

$$\hat{\boldsymbol{\mu}}_5'\hat{\boldsymbol{\Sigma}}^{-1}=(19.9934\quad 69.2849\quad 24.0749\quad 47.9479\quad 575.8803\quad -26.8590)'$$

$$\frac{1}{2}\boldsymbol{\mu}_1'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_1=2\,012.5,\quad \frac{1}{2}\boldsymbol{\mu}_2'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_2=1\,918.4,\quad \frac{1}{2}\boldsymbol{\mu}_3'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_3=1\,869.1,$$

$$\frac{1}{2}\boldsymbol{\mu}_4'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_4=1\,817.8,\quad \frac{1}{2}\boldsymbol{\mu}_5'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_5=1\,670.6$$

利用上述原理，得到下列评价函数：

$$y_5(\boldsymbol{X})=\ln\left(\frac{1}{5}\right)-2\,012.5+24.517X_1+66.958X_2+13.29X_3+81.63X_4+728.65X_5-10.029X_6\qquad(5\text{类})$$

$$y_4(\boldsymbol{X})=\ln\left(\frac{1}{5}\right)-1\,918.4+23.836X_1+67.68X_2+16.116X_3+67.274X_4+700.09X_5-21.12X_6\qquad(4\text{类})$$

$$y_3(\boldsymbol{X})=\ln\left(\frac{1}{5}\right)-1\,869.1+22.944X_1+70.054X_2+18.21X_3+60.741X_4+688.72X_5-26.144X_6\qquad(3\text{类})$$

$$y_2(\boldsymbol{X})=\ln\left(\frac{1}{5}\right)-1\,817.8+21.971X_1+70.74X_2+20.925X_3+56.501X_4+652.88X_5-28.354X_6\qquad(2\text{类})$$

$$y_1(\boldsymbol{X})=\ln\left(\frac{1}{5}\right)-1\,670.6+19.993X_1+69.285X_2+24.075X_3+47.948X_4+575.88X_5-26.859X_6\qquad(1\text{类})$$

用该函数对待测的 3 个矿井进行评价，其结果见表 6.4。

根据待判样本中的数据，分别代入到上面建立的五个判别函数中，每个样本

可以得到五个值，比较这五个值，根据最大值，判定样本属于哪一类。例如对于第一个样本，分别求得各个判别函数的值为：

$y_5^{(1)}=1.6125\times10^3$，$y_4^{(1)}=1.6934\times10^3$，$y_3^{(1)}=1.7393\times10^3$，$y_2^{(1)}=1.7657\times10^3$，$y_1^{(1)}=1.7679\times10^3$，

很明显，几个函数值中 $y_1^{(1)}$ 最大，所以待判样本 1 应该划归为第 1 类。对于待判样本 2 和 3，可类似进行判别，得到表 6.4。

表 6.4　　评价结果

待判样本	矿井编号	X_1	X_2	X_3	X_4	X_5	X_6	期望值
	1	75.58	10.77	28.7	0.88	0.84	1.27	1
	2	85.54	7.12	25.3	1.67	1.15	1.37	3
	3	91.57	5.39	23.1	2.11	1.13	1.58	4

从表 6.4 的结果可以看出，编号为 1 的矿井安全等级为第 1 类，即为不安全等级；编号为 2 的矿井安全等级为第 3 类，即为一般安全等级；编号为 3 的矿井安全等级为第 2 类，即为较不安全等级。其结果与实际符合。

例 6.4　根据植物的症状与受害程度来确定污染类型。假设根据叶色指数 x_1 与植株生长指数 x_2 来区分植物遭受 F、SO_2、HCl 等大气污染物的影响(陈玉成等，1998)。有关训练样本见表 6.5。试根据贝叶斯(Bayes) 判别分析，建立判别函数，并判定另外 3 个待判样本属于哪一类。

表 6.5　　三种大气污染物下的植物反应

组别	序号	叶色指数 x_1	植株生长指数 x_2
第一组遭受 F 污染	1	4.3	15.7
	2	5.6	17.8
	3	4.7	16.9
	4	4.8	16.3
	5	5.3	17.2
	6	4.1	16.0
	7	4.0	15.8
	8	4.6	16.2

续表

组别	序号	叶色指数 x_1	植株生长指数 x_2
第二组 遭受 SO_2 污染	1	9.6	19.6
	2	9.3	19.9
	3	8.7	18.6
	4	8.8	18.9
	5	8.5	19.6
第三组 遭受 HCl 污染	1	10.2	30.3
	2	11.3	28.7
	3	9.8	25.6
	4	7.2	27.6
	5	8.5	29.0
	6	9.6	30.0
待判样本	1	9.2	19.0
	2	4.8	15.3
	3	11.2	30.3

解 根据表中的数据，得到：

$$\hat{\boldsymbol{\mu}}_1 = (4.6750 \quad 16.4875)$$

$$\hat{\boldsymbol{\mu}}_2 = (8.9800 \quad 19.3200)$$

$$\hat{\boldsymbol{\mu}}_3 = (9.4333 \quad 28.5333)$$

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.8198 & 0.3558 \\ 0.3558 & 1.2519 \end{pmatrix}$$

利用 Bayes 原理，建立评价函数为：

$$y_1(\boldsymbol{x}) = \ln\left(\frac{1}{3}\right) - 108.5718 - 0.0153x_1 + 13.1745x_2$$

$$y_2(\boldsymbol{x}) = \ln\left(\frac{1}{3}\right) - 157.5501 + 4.8549x_1 + 14.0529x_2$$

$$y_3(\boldsymbol{x}) = \ln\left(\frac{1}{3}\right) - 326.3905 + 1.8420x_1 + 22.2689x_2$$

根据待判样本中的数据，分别代入到上面建立的三个判别函数中，每个样本可以得到三个值，比较这三个值，根据最大值判定样本属于哪一类。用该函数对待测的待判样本进行评价，其结果如表 6.6。

表 6.6　评价结果

待判样本	样本编号	x_1	x_2	期望值
待	1	9.2	19.0	2
判样	2	4.8	15.3	1
本	3	11.2	30.3	3

从表 6.6 的结果可以看出，编号为 1 的样本属于第二组，即遭受 SO_2 污染；编号为 2 的样本属于第一组，即遭受 F 污染；编号为 3 的样本属于第三组，即遭受 HCl 污染。评价结果与距离判别一致。

例 6.5　仍选取第 4 章第 5 节例 4.6 中 19 个监测站点，对长江流域水质污染进行判别分析。现用 Bayes 方法对长江流域望江楼站污染较重的 4 月份水环境质量进行判别分析。根据水环境质量评价标准在每个等级随机产生 5 个样本，见表 6.7～6.8。

表 6.7　样本数据表

期望值	指标					
	DO(x_1)	高锰酸盐指数(x_2)	BOD_5(x_3)	NH_3-N(x_4)	挥发酚(x_5)	镉(x_6)
1	9.875 3	0.462 3	1.820 5	0.072 9	0.001 8	0.000 8
1	8.641 2	0.037 0	2.464 2	0.066 7	0.001 2	0.000 8
1	9.804 5	1.476 4	0.528 8	0.060 9	0.001 9	0.000 9
1	8.525 7	1.787 3	0.173 7	0.052 9	0.001 6	0.000 0
1	7.847 2	0.405 5	0.596 2	0.090 6	0.000 5	0.000 2
2	6.022 9	3.493 6	3.000 0	0.305 8	0.002 0	0.004 7
2	6.699 0	2.837 3	3.000 0	0.446 2	0.002 0	0.003 1
2	6.304 0	3.344 3	3.000 0	0.443 3	0.002 0	0.001 1
2	7.021 9	2.759 0	3.000 0	0.441 1	0.002 0	0.003 0
2	7.064 2	2.857 8	3.000 0	0.256 6	0.002 0	0.001 8
3	5.193 4	5.364 4	3.302 8	0.770 8	0.002 5	0.003 8
3	5.378 4	5.720 0	3.853 7	0.796 8	0.003 5	0.004 6
3	5.821 6	5.289 8	3.818 0	0.830 1	0.003 0	0.002 2
3	5.341 2	5.068 2	3.727 1	0.654 6	0.004 5	0.003 3
3	5.370 4	5.405 5	3.546 6	0.722 4	0.004 1	0.003 5
4	4.589 6	9.827 4	5.045 2	1.440 1	0.005 9	0.004 9
4	3.542 9	7.009 3	5.751 5	1.368 7	0.005 7	0.001 0
4	4.787 8	6.796 6	4.597 4	1.330 7	0.006 4	0.002 9
4	3.129 6	9.953 3	5.165 6	1.211 7	0.007 6	0.002 3
4	3.865 8	6.903 8	5.159 6	1.380 2	0.007 6	0.003 6

续表

期望值	指标					
	DO(x_1)	高锰酸盐指数(x_2)	BOD_5(x_3)	NH_3-N(x_4)	挥发酚(x_5)	镉(x_6)
5	2.209 1	11.899 1	9.133 3	1.840 4	0.051 5	0.007 8
5	2.794 2	10.295 9	8.411 5	1.525 1	0.047 4	0.006 5
5	2.874 4	10.075 0	9.071 8	1.985 4	0.099 1	0.008 9
5	2.438 7	12.491 6	6.855 9	1.821 7	0.038 8	0.009 8
5	2.726 6	12.059 8	8.978 3	1.634 0	0.049 6	0.009 7

表 6.8　　2000 年 4 月份各站点水环境监测指标实测值

各站点	指标					
	DO(x_1)	高锰酸盐指数(x_2)	BOD_5(x_3)	NH_3-N(x_4)	挥发酚(x_5)	镉(x_6)
攀枝花	7.80	1.80	1.10	0.08	0.000	0.000 0
望江楼	1.40	6.20	24.90	6.22	0.018	0.000 0
高　场	8.00	2.40	0.80	0.08	0.000	0.001 3
朱　沱	9.42	3.04	0.50	0.12	0.000	0.001 5
寸　滩	9.40	2.70	0.60	0.00	0.001	0.001 4
张家界	9.80	1.00	0.60	0.06	0.000	0.000 0
吉　首	7.60	2.50	5.60	0.22	0.000	0.000 0
芷　江	6.60	1.50	0.90	0.32	0.000	0.000 0
坝　上	7.60	3.90	3.00	1.00	0.000	0.000 0
津　市	8.00	1.90	0.90	0.22	0.002	0.000 0
石　门	9.00	1.30	0.60	0.22	0.002	0.000 0
益　阳	8.30	1.70	0.20	0.09	0.000	0.000 0
湘　潭	9.10	2.50	2.00	0.30	0.000	0.000 0
株　洲	7.30	3.80	1.60	0.22	0.000	0.000 0
衡　阳	8.00	3.10	1.50	0.34	0.000	0.000 0
长　沙	8.10	2.30	0.50	0.17	0.000	0.002 7
吉　安	8.70	2.50	2.70	0.21	0.000	0.000 0
中　山	10.10	1.70	1.50	0.10	0.000	0.000 0
宣　城	9.30	1.50	0.60	0.00	0.000	0.000 0

解　根据表中的数据，得到样本均值：

$$\hat{\boldsymbol{\mu}}_1=(8.938\,8\quad 0.833\,7\quad 1.116\,7\quad 0.068\,8\quad 0.001\,4\quad 0.000\,5)$$

$$\hat{\boldsymbol{\mu}}_2=(6.622\,4\quad 3.058\,4\quad 3.000\,0\quad 0.378\,6\quad 0.002\,0\quad 0.002\,7)$$

$$\hat{\boldsymbol{\mu}}_3=(5.421\,0\quad 5.369\,6\quad 3.649\,6\quad 0.754\,9\quad 0.003\,5\quad 0.003\,5)$$

$\hat{\boldsymbol{\mu}}_4=(3.9831 \quad 8.0981 \quad 5.1439 \quad 1.3463 \quad 0.0066 \quad 0.0029)$

$\hat{\boldsymbol{\mu}}_5=(2.6086 \quad 11.3643 \quad 8.4902 \quad 1.7613 \quad 0.0573 \quad 0.0085)$

协方差：

$$\hat{\boldsymbol{\Sigma}}=\begin{pmatrix} 0.3190 & -0.0954 & 0.0137 & 0.0059 & 0.0007 & 0.0001 \\ -0.0954 & 0.9266 & -0.2140 & -0.0080 & -0.0036 & 0.0004 \\ 0.0137 & -0.2140 & 0.4185 & 0.0035 & 0.0024 & -0.0001 \\ 0.0059 & -0.0080 & 0.0035 & 0.0107 & 0.0005 & 0.0000 \\ 0.0007 & -0.0036 & 0.0024 & 0.0005 & 0.0001 & 0.0000 \\ 0.0001 & 0.0004 & -0.0001 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

利用 Bayes 原理，所建立的评价函数为：

$$y_1(\boldsymbol{x})=\ln\left(\frac{1}{5}\right)-141.6957+30.7x_1+6.1x_2+4.5x_3+6.1x_4-102.2x_5-3209.7x_6$$

$$y_2(\boldsymbol{x})=\ln\left(\frac{1}{5}\right)-112.879+23.2x_1+8.3x_2+11.7x_3+46.4x_4-322.4x_5-1928.5x_6$$

$$y_3(\boldsymbol{x})=\ln\left(\frac{1}{5}\right)-142.9039+19.9x_1+11.4x_2+15.3x_3+95.3x_4-495.4x_5-2683.3x_6$$

$$y_4(\boldsymbol{x})=\ln\left(\frac{1}{5}\right)-262.3979+16.5x_1+16.0x_2+21.9x_3+176.5x_4-811.0x_5-5185.7x_6$$

$$y_5(\boldsymbol{x})=\ln\left(\frac{1}{5}\right)-418.1121+11.3x_1+21.4x_2+31.4x_3+193.0x_4-456.6x_5-1976.9x_6$$

根据待判样本中的数据，分别代入到上面建立的五个判别函数中，每个样本可以得到五个值，比较这五个值，根据最大值判定样本属于哪一类。用该函数对待测的 19 个监测站点的水质污染进行判别分析，其评价结果见表 6.9。

表 6.9　　评价结果

各站点	判别函数值（$\times10^3$）					期望值
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	
攀枝花	0.112 8	0.098 0	0.055 7	−0.068 6	−0.243 2	1
望江楼	0.086 2	0.542 9	0.918 7	1.485 7	1.702 8	5
高　场	0.117 1	0.101 7	0.058 4	−0.069 0	−0.240 1	1

续表

各站点	判别函数值（×10³）					期望值
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	
朱　沱	0.162 9	0.137 9	0.092 7	−0.035 9	−0.212 5	1
寸　滩	0.160 2	0.130 1	0.078 3	−0.060 9	−0.240 3	1
张家界	0.167 1	0.131 0	0.076 8	−0.063 0	−0.257 4	1
吉　首	0.132 2	0.158 2	0.141 8	0.062 4	−0.062 2	2
芷　江	0.074 7	0.076 5	0.048 2	−0.055 2	−0.223 2	2
坝　上	0.133 7	0.175 7	0.192 4	0.165 7	0.036 7	3
津　市	0.119 3	0.107 0	0.070 1	−0.045 0	−0.219 0	1
石　门	0.145 1	0.121 7	0.078 6	−0.044 7	−0.230 0	1
益　阳	0.123 6	0.098 7	0.051 7	−0.079 9	−0.266 1	1
湘　潭	0.162 5	0.154 7	0.124 3	0.022 5	−0.142 9	1
株　洲	0.112 8	0.115 4	0.089 5	−0.009 1	−0.163 3	2
衡　阳	0.130 3	0.130 2	0.105 4	0.010 2	−0.150 4	1
长　沙	0.114 3	0.101 1	0.059 5	−0.066 9	−0.236 0	1
吉　安	0.152 8	0.149 4	0.118 4	0.015 4	−0.142 8	1
中　山	0.184 9	0.156 1	0.108 4	−0.020 1	−0.203 0	1
宣　城	0.154 4	0.120 8	0.066 9	−0.073 8	−0.263 9	1

【思考题 6】

1. 试述距离判别的基本步骤。
2. 试述 Fisher 判别的基本步骤。
3. 试述 Bayes 判别的基本步骤。
4. 为取得评价所需的监测数据，在项目及周边地区布设了 5 个空气监测点，分别记为 1，2，3，4，5 号点，按要求在各点对 SO_2、NOx、TSP 三种因子进行监测，得到三种评价因子的日平均浓度值，见表 6.10。

表 6.10　　各监测点的监测值表　单位：mg/m^3

测点号	日平均值		
	SO_2	NOx	TSP
1	0.006 0	0.022	0.327
2	0.009 0	0.018	0.190
3	0.010 5	0.015	0.253
4	0.004 0	0.025	0.160
5	0.003 0	0.018	0.053

评价选用环境空气质量评价标准，该标准的浓度限值见表 6.11，该标准体系共分三级，对应三类环境功能区。一类区为自然保护区、风景名胜区和其他需要特殊保护的地区；二类区为城镇规划中确定的居住区、商业交通居民混合区、文化区、一般工业区和农村地区；三类区为特定工业区。并规定一类区执行一级标准；二类区执行二级标准；三类区执行三级标准。该项目为房地产开发项目，建成后为高级住宅区，项目所在地为大城市的郊区。试用距离判别法判断各监测点所处环境功能区的类型，这里每一级标准取 5 个样本，并给出判别函数。

表 6.11　　大气环境质量标准表(GB 3095—1996)　单位：mg/m^3

污染物名称	日平均浓度限值		
	一级标准	二级标准	三级标准
SO_2	0.05	0.15	0.25
NOx	0.10	0.10	0.15
TSP	0.12	0.30	0.50

5. 为了了解某一河流 As、Pb 的污染状况，分别在甲、乙两地监测，采样分析得水中的 As、Pb 浓度与底泥中的 As、Pb 浓度。现有两个未知样本 A，B，相应的监测数据一并列入表 6.12，试用 Fisher 判别方法判断两个未知样本是从甲、乙两个区域中的哪一个采得的，并给出判别函数。

表 6.12 河流 As、Pb 的污染监测表

地区	样本序号	水中 As /(mg・L^{-1})	泥中 As /(mg・kg^{-1})	水中 Pb /(mg・L^{-1})	泥中 Pb /(mg・kg^{-1})
甲地	1	4.67	22.31	12.31	47.80
	2	4.63	28.82	16.18	62.55
	3	3.54	15.29	7.58	43.20
乙地	1	1.06	2.18	1.22	20.60
	2	0.80	3.85	4.06	47.10
	3	0.00	11.40	3.50	0.00
	4	2.42	3.66	2.14	15.00
样本 A		2.79	13.85	7.80	49.60
样本 B		2.40	7.90	4.30	33.20

6. 试述距离判别、Fisher 判别、Bayes 判别的区别与联系。

【参考文献】

[1] 陈玉成，吕宗清，李章平. 环境数学分析 [M]. 重庆：西南师范大学出版社，1998.

[2] 何晓群. 现代统计分析方法与应用 [M]. 北京：中国人民大学出版社，2003.

[3] 雷兢，沈斐敏. 贝叶斯(Bayes)判别分析理论在安全评价中的应用 [J]. 工业安全与环保，2004，30(5)：39-40.

第7章 环境主成分分析

在环境统计学中，经常会遇到环境因素复杂、因子众多的环境数据的处理和分析。如何对这些因子提取主要成分，以便对实际环境问题进行系统分析、评价是很重要的。主成分分析便是处理这类问题的有效工具之一。

主成分分析(principal components analysis)也称主分量分析，是由 Hotelling 在 1933 年首先提出的。目前，主成分分析在经济管理、环境科学与环境工程等许多方面都有广阔的应用前景。主成分分析是利用对高维变量空间进行降维处理的思想，把多个指标转化为少数几个综合指标的多元统计分析方法。

本章的主要内容是：

- 主成分分析概述；
- 主成分分析计算原理；
- 主成分分析基本性质；
- 环境应用。

7.1 主成分分析概述

主成分分析的工作对象是一张样本点乘以变量指标的数据表。它的工作目的就是在保证信息损失量最小的前提下，尽可能提取问题的主要方面，从而对多变量数据进行最佳综合简化。如果在原数据表中有 p 个变量 x_1，x_2，…，x_p，主成分分析法就是对这个数据表中的信息进行重新调整组合，从中提取 m 个综合变量 F_1，F_2，…，$F_m(m<p)$，使这 m 个综合变量能最多地概括原数据表中的信息。也就是说，主成分分析可以在力保数据信息损失最少的原则下，对高维变量空间进行降维处理。很显然，在一个低维空间作系统分析要比在高维空间容易得多。英国统计学家斯格特(M. Scott)在 1961 年对 157 个英国城镇的发展水平进行调查时，测量的原始变量有 57 个。而通过主成分分析发现，只需 5 个新的综合变量(它们是原变量的线性组合)就可以用 95%的精度概括原数据表中的信息。这样，问题的研究一下子从 57 维降到 5 维。可以想象，在 5 维空间对系统进行任何分析都比在 57 维中更加方便、快捷。

在对多变量系统进行综合简化时，有一种情况尤其引起人们的关注：如果能将一个 p 维变量系统有效地降至 2 维，就可以在一个平面图上描绘出每一个样本

点，从而直接观察样本点之间的相似关系以及样本点群的分布特点和结构。所以，主成分分析使高维空间中数据点的可视性成为可能。在数据信息的分析过程中，对直观图像的观察是一种重要的分析手段，它可以更好地协助系统分析人员进行思考与判断，及时发现大规模数据群中的普遍规律与特殊现象，提高数据信息的分析效率。

近年来，随着多元统计方法在环境、经济和管理等领域的推广与普及，主成分分析又有了十分重要的应用，从而成为构造系统评估指数、对系统中的元素进行评估排序的常用方法之一。事实上，如果能在 p 维变量的数据表中有效地提取一个综合变量，而这个综合变量能以较高的精度概括原数据表中的信息，它就有可能成为一个系统评估指数。用主成分分析构造评估指数的方法有许多成功的应用实例。英国统计学家肯道尔(M. Kendall)曾对48个郡的小麦、大麦、燕麦、土豆、菜豆、马铃薯、萝卜、饲料甜菜、临时牧场干草、永久牧场干草10种主要农作物进行生产调查。在进行主成分分析后，以47.6%的精度提取了一个最佳的综合变量(第一主成分) F_1 作为系统评估指数。肯道尔将这个综合变量 F_1 称为生产能力水平，并利用这一指数，把英国各地区农作物生产按 F_1 排序和分类。而这一评估结果与当时有关农业生产能力地理分布的实际情况是十分一致的。可见，主成分分析作为多元统计分析方法，在实践中具有重要的研究意义。

7.2 主成分分析计算原理

在主成分分析中，在数据表内提取的综合变量被称为主成分。怎样在数据表中提取主成分，使之能最好地概括原数据表中的信息？又怎样能够将一个高维空间进行降维处理？

在统计学中，说到数据集合中的信息，一般常常指这个集合中的数据变异的情况。例如，在回归分析中，回归方程的测定系数就是要测量在多大程度上能用回归方程来解释 y 的变异。而在一张数据表中，数据集合的变异信息可以用全部变量方差的总和来测量。方差越大，数据中包含的信息就越多。

假设有一个二维数据表($p=2$)，表中样本点的分布如图7-1所示，呈圆棍形状，重心是 g。很显然，在沿棍子轴的方向 a_1 上，数据的离差最大，因此，所反映的数据信息也最多，这个方向被称为数据变异最大的方向。如果将坐标原点平移到 g，并且作旋转变换，得到一个正交坐标系 a_1ga_2。将样本点在 a_1 轴上投影得到新变量 F_1，则 F_1 是一个能携带最多原变异信息的综合变量，这就是所要提取的第一主成分。而如果省略 a_2 轴，就会得到一个简化的一维数据系统。所以，对高维数据系统进行降维处理的核心思想，就是省却变异不大的变量方向

（何晓群，2003）。

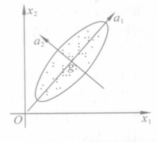


图 7-1　二维数据图

又如，一个三维数据表的样本点分布是球形的。假若这个球是饼状的，其变异较大的方向为 a_1，a_2，而 a_3 方向上的变异很小，即在该方向上样本点取值没有很大的差别，就可以不考虑 a_3 方向。若以 a_1ga_2 作为新的坐标系来分析数据，则原三维空间的数据点就可以在二维平面上得以显示。将所有样本点分别在 a_1，a_2 上投影，就得到携带原数据变异信息最多的新综合变量 F_1，以及携带原变异信息次多的新综合变量 F_2。

推广到更一般的情形。设原始数据表中的变量为 x_1，x_2，…，x_p。主成分分析的过程实质上是对原坐标系进行平移和旋转变换，使得新坐标系的原点与样本点集合的重心重合，新坐标系的第一轴与数据变异的最大方向对应，新坐标系的第二轴与第一轴标准正交，并且对应于数据变异的第二大方向，依次类推。这些新轴分别被称为第一主轴，第二主轴，……。若经舍弃少量信息后，由主轴 a_1，a_2，…，a_m 构成的子空间能够十分有效地表示原数据的变异情况，则原来的 p 维空间就被降至 m 维。这个新生成的 m 维子空间被称为 m 维主超平面。当 $m=2$ 时，就称其为主平面。可以用原样本点集合在主超平面的第 h 主轴上的投影构成综合变量 $F_h \in R^n (h=1, 2, \cdots, m)$，称为第 h 主成分。若以方差 $D(F_h)$ 度量第 h 主成分 F_h 所携带的变异信息，则主成分分析的结果是：

$$D(F_1) \geqslant D(F_2) \geqslant \cdots \geqslant D(F_m) > 0$$

记 $\boldsymbol{X}$ 是一个有 n 个样本点和 p 个变量的数据表：

$$\boldsymbol{X} = (x_{ij})_{n \times p} = \begin{bmatrix} \boldsymbol{e}'_1 \\ \vdots \\ \boldsymbol{e}'_n \end{bmatrix}$$

样本点 $\boldsymbol{e}_i = (x_{i1}, \cdots, x_{ip})' \in R^p$。

为推导方便，且不失一般性，设该数据表是标准化的（即 $E(x_j)=0$，

$D(x_j)=1$)。现要求一个综合变量 F_1，F_1 是 x_1，x_2，…，x_p 线性组合，记 $\boldsymbol{a}_1=(a_{11}, a_{12}, \cdots, a_{1p})'$，$\boldsymbol{x}=(x_1, x_2, \cdots, x_p)'$，也即：

$$F_1=\boldsymbol{a}_1'\boldsymbol{x},\quad \|\boldsymbol{a}_1\|=1$$

要使得 F_1 能携带最多的原变异信息，即要求 F_1 的方差取到最大值。这里，我们不限定样本点集合一定是随机抽样得到的，因此，F_1 的方差为：

$$D(F_1)=\frac{1}{n}\boldsymbol{a}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_1=\boldsymbol{a}_1'\boldsymbol{V}\boldsymbol{a}_1$$

这里，记 $\boldsymbol{V}=\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ 是 $\boldsymbol{X}$ 数据表的协方差矩阵。当 $\boldsymbol{X}$ 中的变量均是标准化变量时，$\boldsymbol{V}$ 就是 $\boldsymbol{X}$ 的相关系数矩阵。

把上面的问题写成数学表达式，即求优化问题：

$$\max_{\|\boldsymbol{a}_1\|=1} \boldsymbol{a}_1'\boldsymbol{V}\boldsymbol{a}_1$$

采用拉格朗日(Lagrange)算法求解，记 λ_1 是拉格朗日系数，令：

$$L=\boldsymbol{a}_1'\boldsymbol{V}\boldsymbol{a}_1-\lambda_1(\boldsymbol{a}_1'\boldsymbol{a}_1-1)$$

对 L 分别求关于 $\boldsymbol{a}_1$ 和 λ_1 的偏导，并令其为零，有：

$$\frac{\partial L}{\partial \boldsymbol{a}_1}=2\boldsymbol{V}\boldsymbol{a}_1-2\lambda_1\boldsymbol{a}_1=0 \tag{7.1}$$

$$\frac{\partial L}{\partial \lambda_1}=-(\boldsymbol{a}_1'\boldsymbol{a}_1-1)=0 \tag{7.2}$$

由式(7.1)得：

$$\boldsymbol{V}\boldsymbol{a}_1=\lambda_1\boldsymbol{a}_1 \tag{7.3}$$

由此可知，$\boldsymbol{a}_1$ 是 $\boldsymbol{V}$ 的一个标准化特征向量，它所对应的特征值是 λ_1。而根据目标函数及上式，有：

$$D(F_1)=\boldsymbol{a}_1'\boldsymbol{V}\boldsymbol{a}_1=\boldsymbol{a}_1'(\lambda_1\boldsymbol{a}_1)=\lambda_1\boldsymbol{a}_1'\boldsymbol{a}_1=\lambda_1 \tag{7.4}$$

所以，$\boldsymbol{a}_1$ 所对应的特征值 λ_1 应取到最大值。

换句话说，我们所要求的 $\boldsymbol{a}_1$ 是矩阵 $\boldsymbol{V}$ 的最大特征值 λ_1 所对应的标准化特征向量。这里，$\boldsymbol{a}_1$ 被称为第一主轴，$F_1=\boldsymbol{a}_1'\boldsymbol{X}$ 被称为第一主成分。

接着，可以求第二主轴 $\boldsymbol{a}_2$，$\boldsymbol{a}_2$ 与 $\boldsymbol{a}_1$ 标准正交($\boldsymbol{a}_2'\boldsymbol{a}_1=0$，$\|\boldsymbol{a}_2\|^2=1$)，并且仅次于第一主成分 F_1，第二主成分 $F_2=\boldsymbol{a}_2'\boldsymbol{x}$ 是携带变异信息第二大的成分。F_2 的方差为：

$$D(F_2)=\frac{1}{n}\boldsymbol{a}_2'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_2=\boldsymbol{a}_2'\boldsymbol{V}\boldsymbol{a}_2$$

写成优化问题，即：

$$\max \boldsymbol{a}_2'\boldsymbol{V}\boldsymbol{a}_2$$

$$\boldsymbol{a}_2'\boldsymbol{a}_1=0,\quad \boldsymbol{a}_2'\boldsymbol{a}_2=1$$

类似于求 F_1 的过程，定义拉格朗日函数为：

$$L=\boldsymbol{a}_2'\boldsymbol{V}\boldsymbol{a}_2-\lambda_2(\boldsymbol{a}_2'\boldsymbol{a}_2-1)$$

求 L 关于 $\boldsymbol{a}_2$ 与 λ_2 的偏导，并令之为零，得：

$$\boldsymbol{V}\boldsymbol{a}_2=\lambda_2\boldsymbol{a}_2 \tag{7.5}$$

$$\boldsymbol{a}_2'\boldsymbol{a}_2=1 \tag{7.6}$$

$\boldsymbol{a}_2$ 是矩阵 $\boldsymbol{V}$ 的标准化特征向量，它所对应的特征根是 λ_2，而

$$\lambda_2=\boldsymbol{a}_2'\boldsymbol{V}\boldsymbol{a}_2=D(F_2) \tag{7.7}$$

由于有约束 $\boldsymbol{a}_2'\boldsymbol{a}_1=0$，因此，这时 λ_2 只能是矩阵 $\boldsymbol{V}$ 的第二大特征值，$\boldsymbol{a}_2$ 是对应于 $\boldsymbol{V}$ 第二大特征值的标准化特征向量。

依次类推，可求得 $\boldsymbol{X}$ 数据表的第 h 主轴 a_h，它是协方差矩阵 $\boldsymbol{V}$ 的第 h 个特征值 λ_h 所对应的标准化特征向量。而第 h 主成分 F_h：

$$F_h=\boldsymbol{a}_h'\boldsymbol{x} \tag{7.8}$$

由：

$$D(F_h)=\boldsymbol{a}_h'\boldsymbol{V}\boldsymbol{a}_h=\boldsymbol{a}_h'(\lambda_h\boldsymbol{a}_h)=\lambda_h \tag{7.9}$$

因此，有 $D(F_1)\geqslant D(F_2)\geqslant\cdots\geqslant D(F_m)$。

所以，用数据变异大小来反映数据中的信息，则第一主成分 F_1 携带的信息量最大，F_2 次之，以此类推。如果抽取了 m 个主成分，这 m 个主成分所携带的信息量总和为：

$$\sum_{h=1}^{m}D(F_h)=\sum_{h=1}^{m}\lambda_h \tag{7.10}$$

归纳上述分析可以看出，主成分分析的计算步骤如下：

(1)对数据进行标准化处理：

$$\tilde{x}_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}\quad(i=1,2,\cdots,n;\ j=1,2,\cdots,p)$$

式中，$\bar{x}_j$ 是 x_j 的样本均值；s_j 是 x_j 的样本标准差。

标准化处理的目的是使样本点集合的重心与坐标原点重合，而压缩处理可以消除由量纲不同所引起的虚假变异信息，使分析结果更加合理。为方便起见，仍记标准化处理的矩阵为 $\boldsymbol{X}$。

(2)计算标准化数据矩阵 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{V}$，这时 $\boldsymbol{V}$ 又是 $\boldsymbol{X}$ 的相关系数矩阵；

(3)求 $\boldsymbol{V}$ 的前 m 个特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$，以及对应的特征向量 $\boldsymbol{a}_1$，$\boldsymbol{a}_2$，…，$\boldsymbol{a}_m$，要求它们是标准正交的；

(4)求第 h 个成分的累计贡献率：

$$\eta_h=\frac{\sum_{i=1}^{h}\lambda_i}{\sum_{i=1}^{m}\lambda_i}\qquad (m\geqslant h)$$

(5)求第 h 主成分 F_h，有：

$$F_h=\boldsymbol{a}_h'\boldsymbol{x}=\sum_{j=1}^{p}a_{hj}x_j \tag{7.11}$$

式中，a_{hj} 是主轴 a_h 的第 j 个分量。所以，主成分 F_h 是原变量 x_1，x_2，…，x_p 的线性组合，组合系数恰好为 a_{hj}。从这个角度，又可以说 F_h 是一个新的综合变量。

7.3 主成分分析的性质

主成分分析主要有以下几条基本性质：

(1)主成分 F_h 的样本均值等于零。记 $E(F_h)$ 为样本均值，有：

$$E(F_h)=\frac{1}{n}\sum_{i=1}^{n}F_h(i)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}a_{hj}x_{ij}=\sum_{j=1}^{p}a_{hj}\left(\frac{1}{n}\sum_{i=1}^{n}x_{ij}\right)=0$$

式中，$F_h(i)$ 是 F_h 的第 i 个分量。

(2)F_h 的样本方差等于 λ_h，即：

$$D(F_h)=\lambda_h$$

这个结论在式(7.9)中已经给予证明。

(3)各个主成分之间是互不相关的，即样本协方差为：

$$\mathrm{Cov}(F_h,\ F_l)=0\qquad (\forall\, l\neq h) \tag{7.12}$$

证明：

$$\mathrm{Cov}(F_h,\ F_l)=\boldsymbol{a}_h'\boldsymbol{V}\boldsymbol{a}_l=\boldsymbol{a}_h'(\lambda_l\boldsymbol{a}_l)=\lambda_l\,\boldsymbol{a}_h'\boldsymbol{a}_l=0$$

这个性质说明经过主成分分析，可将原始测量的 p 个相关变量变换成一组相互无关的正交变量(即主成分之间的协方差等于零)。在许多实际应用中，变量系统的正交性是十分有益的性质。由于各个变量中所含的信息是互补的，并且在信息中间没有交叉重叠，这将对进一步开展其他方面的统计分析带来许多便利。

从上面的讨论可知，主成分分析过程可示意为：

$$(x_1,\ \cdots,\ x_p)\xrightarrow{\text{主成分分析}}(F_1,\ \cdots,\ F_m)\quad (m<p)$$

记 $\boldsymbol{F}=(F_1,\ \cdots,\ F_m)$ 是新变量系统下的数据表，它是由原数据表 $\boldsymbol{X}=(x_1,\ \cdots,\ x_p)$ 经数学变换，并省略一部分信息而得到的。可以记为：

$$F=(F_1,\ \cdots,\ F_m)=\begin{pmatrix} e_1' \\ \vdots \\ e_n' \end{pmatrix}$$

$\hat{e}_i$ 是数据表 F 的第 i 个样本点，它为：

$$\hat{e}_i=(F_1(i),\ \cdots,\ F_m(i))' \quad (i=1,\ 2,\ \cdots,\ n)$$

记集合 $A=\{\hat{e}_i\}$ $(i=1,\ 2,\ \cdots,\ n)$，则 A 的重心是与原点重合的，即：

$$\hat{\boldsymbol{g}}=\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i=\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}F_1(i) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}F_m(i) \end{pmatrix}=\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

因此，$\hat{e}_i$ 到重心 $\hat{\boldsymbol{g}}=\boldsymbol{0}$ 的距离为：

$$d^2(\hat{e}_i,\ \boldsymbol{0})=\sum_{h=1}^{m}F_h^2(i) \tag{7.13}$$

这里特别需要指出的是，如果取 $m=p$，则主成分分析只相当于在原 p 维空间中作了一次坐标变换，而没有任何信息损失。这时，样本点 $\boldsymbol{e}_i=(F_1(i),\ \cdots,\ F_p(i))'\in R^p$ 不过是原样本点 $\boldsymbol{e}_i$ 在新坐标系下的重新表示。所以，这时 $\boldsymbol{e}_i$ 到原点的距离为：

$$d^2(\boldsymbol{e}_i,\ \boldsymbol{0})=\sum_{h=1}^{m}F_h^2(i) \tag{7.14}$$

7.4 环境应用

由于环境系统是一个复杂的开放大系统，环境质量的变化是各变量(因素)综合作用的结果，主成分分析的重要作用之一就是区分主要因素与次要因素。除此之外，主成分还有下列一些用途(陈玉成等，1998)：

(1)压缩原始数据，减轻环境工作者综合分析的负担；

(2)使用综合评价值对样本和变量进行分类，探索污染源，分析污染物的时空分布规律；

(3)确定环境质量评价中各要素的相对重要性(权重)；

(4)分析环境污染的理化过程；

(5)环境质量监测的优化布点；

(6)与回归分析、聚类分析、因子分析等其他多元统计分析结合，从原始数

据中提取更多的有用信息。

在进行主成分分析时，需注意以下几点：

(1)只取第一主成分。一般说来，第一主成分在较大程度上全面综合了各指标的信息，用它就可作综合评价指标。

(2)当原始各指标对综合指标都为正指标时，主成分 F_h 中系数绝对值较大者应有同一的符号且都为正值。若系数绝对值较大者都为负值(即这个主成分与这些原指标为负相关)，此时应把相应特征向量改向。

(3)若觉得用一个主成分解释的方差不够大，综合程度不够，而用多个主成分综合又不合适时，用因子分析中方差最大的正交旋转可能会取得较好的效果。

例 7.1 主成分分析在大气环境质量评价中的应用

将大气环境质量的评价等级从好到坏分为4级，大气环境质量评价标准见表7.1。1993年西南铝加工厂各季度单元中各评价指标的实际监测浓度值见表7.2。试对各季度大气环境质量进行评价。

表 7.1　大气环境质量评价标准　单位：mg/m^3

大气环境质量级别	指标		
	$SO_2(x_1)$	$NO_x(x_2)$	$TPS(x_3)$
Ⅰ级(e_1)	0.05	0.05	0.15
Ⅱ级(e_2)	0.15	0.10	0.30
Ⅲ级(e_3)	0.25	0.15	0.50
Ⅳ级(e_4)	0.85	0.50	1.70

注：Ⅳ为严重污染临界浓度。

表 7.2　各季度单元中各评价指标实测浓度值　单位：mg/m^3

季度单元	指标		
	$SO_2(x_1)$	$NO_x(x_2)$	$TPS(x_3)$
第一季度(e_5)	0.046	0.036	0.086
第二季度(e_6)	0.139	0.044	0.152
第三季度(e_7)	0.032	0.014	0.159
第四季度(e_8)	0.056	0.016	0.183

解　原决策矩阵：

$$X=\begin{pmatrix} 0.0500 & 0.0500 & 0.1500 \\ 0.1500 & 0.1000 & 0.3000 \\ 0.2500 & 0.1500 & 0.5000 \\ 0.8500 & 0.5000 & 1.7000 \\ 0.0460 & 0.0360 & 0.0860 \\ 0.1390 & 0.0440 & 0.1520 \\ 0.0320 & 0.0140 & 0.1590 \\ 0.0560 & 0.0160 & 0.1830 \end{pmatrix}$$

1. 标准差标准化处理后的矩阵为：

$$A=\begin{pmatrix} -0.5347 & -0.3921 & -0.4703 \\ -0.1700 & -0.0845 & -0.1923 \\ 0.1947 & 0.2229 & 0.1784 \\ 2.3828 & 2.3748 & 2.4025 \\ -0.5493 & -0.4780 & -0.5889 \\ -0.2102 & -0.4289 & -0.4666 \\ -0.6004 & -0.6133 & -0.4536 \\ -0.5129 & -0.6010 & -0.4092 \end{pmatrix}$$

2. 计算标准化数据矩阵 $\boldsymbol{A}$ 的协方差矩阵：

$$C=\begin{pmatrix} 1.0000 & 0.9936 & 0.9925 \\ 0.9936 & 1.0000 & 0.9931 \\ 0.9925 & 0.9931 & 1.0000 \end{pmatrix}$$

3. 求 $\boldsymbol{C}$ 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$ 以及对应的特征向量 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$，$\boldsymbol{u}_3$，要求它们是标准正交的。

$$U=\begin{pmatrix} 0.3333 & 0.3771 & 0.3551 \\ 0.3334 & 0.1229 & -0.4999 \\ 0.3333 & -0.5001 & 0.1450 \end{pmatrix} = (\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \boldsymbol{u}_3)$$

$$\lambda_1 = 2.9862 \quad \lambda_2 = 0.0076 \quad \lambda_3 = 0.0063$$

4. 累计贡献率：　$\alpha_1 = 0.9954 > 85\%$

5. 求第一主成分 F_1，有：

$$F_1 = \boldsymbol{A}\boldsymbol{u}_1 = (-0.4657\ \ -0.1490\ \ 0.1986\ \ 2.3867\ \ -0.5388\ \ -0.3685\ \ -0.5558\ \ -0.5077)$$

6. 综合评价：

$$F_1=(-0.4657\quad -0.1490\quad 0.1986\quad 2.3867\quad -0.5388\quad -0.3685\quad -0.5558\quad -0.5077)$$

按 F_1 由小到大的顺序排列方案的优先次序，结果是：

$$e_7<e_5<e_8<e_1<e_6<e_2<e_3<e_4$$

根据上面的计算结果，可判断西南铝加工厂各季度大气环境质量从优到劣依次为：第三季度、第一季度、第四季度、第二季度。由于第三季度、第一季度、第四季度的大气环境质量优于Ⅰ级标准最低界限值，所以它们都属于Ⅰ级，而第二季度的大气环境质量介于Ⅰ级和Ⅱ级标准最低界限值之间，所以它属于Ⅱ级。本节方法与污染损失率法、模糊综合评判方法、灰关联分析方法、多目标决策—理想区间法和变权识别模型等的评价结果相一致。各评价方法的对比结果见表7.3。

表7.3　主成分分析法评价结果

评价方法	第一季度	第二季度	第三季度	第四季度
主成分分析法	Ⅰ	Ⅱ	Ⅰ	Ⅰ

例7.2　主成分分析在天然气开发环境影响评价中的应用

天然气开发钻井施工过程中，环境影响主要有以下几个方面：①地震勘探作业产生的爆炸噪声将影响环境及其附近的居民和野生动物；②钻井作业对环境的影响因素主要有修建钻井井场和井场公路两个方面，平均每口井需征用井场用地面积约为 6.67×10^4 m^2，井场公路的占地面积取决于井场与交通公路的距离；③道路施工会造成季节性的水土流失问题，在修建集气管道时，敷设管沟的开挖施工将破坏管道沿线两侧各7.5 m范围内的植被，在施工期会对坡度大于5°的施工地段的水土保持产生较大的影响；④天然气开发正常运行期间对环境的影响较小，主要是钻井过程中柴油机会产生废气，钻井、起下钻、固井作业等产生废水，机械设备运转会产生噪声，此外，还有钻井岩屑、废泥浆等产生；⑤采气生产活动中当集气设施(如管道、分离器、阀门等)需要检修或出现爆管事故时将进行天然气燃烧放空作业，放空产生的热辐射和噪声将对其周围的居民和植被产生一定影响；⑥采气生产后期，地层水含盐量很高，外排时会对受纳水体产生一定影响；⑦天然气净化时，含硫天然气燃烧后排放的 SO_2 会污染大气环境，天然气净化厂还将排放少量废水和产生轻微噪声污染(周晓东等，2000；师春元等，2001)。

四川盆地是我国天然气开采较早、储量较丰富的区域，在满足四川省和重庆直辖市需求的同时，通过管道外送部分剩余气量。德阳地区孝泉—新场—合兴场气田开发工程(以下简称孝新合气田工程)位于成都平原西侧，在德阳市北侧，跨德阳市所辖绵竹县及德阳市旌阳区的众多乡镇，距成都约 80 km，地理坐标为东经 104°11′～104°34′，北纬 31°08′～31°20′。其区域地理位置见图 7-2。孝新合气田工程所在地区地表主要河流有石亭江、绵远河、凯江三条。

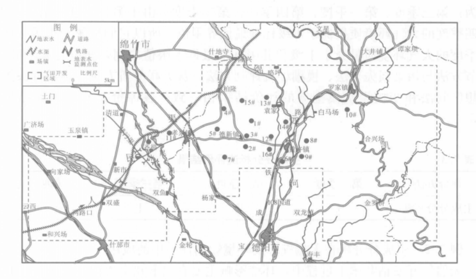


图 7-2　德阳地区孝泉—新场—合兴场气田开发工程范围示意图

在区域开发区内 3 条主要河流上选择多个具有代表性的地表水监测断面，根据项目特点和实际需求，主要的测量指标有 pH 值、COD_{Mn}、石油类、挥发酚四个指标。天然气开发前、后均在河流上同一个采样点按每天 7 个时段进行取样分析，其监测数据如表 7.4 所示。将每一流域上不同时段的监测数据取平均值，可得代表每一流域的水样监测分析结果，具体见表 7.5。

表 7.4 天然气开发前、后流域地表水水质监测数据

单位：mg/L(pH值除外)

时段	河流	监测项目	8：00	10：00	12：00	14：00	16：00	18：00	20：00
开发前(2000年5月6日)	绵远河	pH值	7.50	7.60	7.30	7.82	8.00	7.10	7.60
		COD_{Mn}	15.00	14.10	13.60	14.60	14.20	14.00	15.80
		石油类	0.200	0.250	0.260	0.190	0.210	0.240	0.270
		挥发酚	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	凯江	pH值	9.30	9.60	8.75	8.90	9.00	9.20	10.00
		COD_{Mn}	28.61	29.20	28.90	27.90	29.20	29.10	29.32
		石油类	0.108	0.130	0.126	0.120	0.128	0.132	0.131
		挥发酚	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	石亭江	pH值	8.00	8.20	8.50	8.03	7.80	7.90	8.03
		COD_{Mn}	23.00	22.60	21.90	22.60	22.30	21.80	23.16
		石油类	0.108	0.125	0.127	0.120	0.106	0.130	0.124
		挥发酚	0.001	0.001	0.001	0.001	0.001	0.001	0.001
开发后(2005年4月8日)	绵远河	pH值	8.13	8.18	8.24	8.09	8.06	8.21	8.23
		COD_{Mn}	11.00	10.20	10.06	10.03	10.30	10.31	10.69
		石油类	0.060	0.073	0.076	0.068	0.070	0.069	0.081
		挥发酚	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	凯江	pH值	7.20	7.36	7.26	7.33	7.28	7.38	7.38
		COD_{Mn}	21.20	20.10	20.06	19.87	20.30	19.92	19.04
		石油类	0.062	0.068	0.053	0.049	0.051	0.060	0.070
		挥发酚	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	石亭江	pH值	7.60	7.53	7.47	7.32	7.28	7.26	7.41
		COD_{Mn}	31.00	30.04	30.86	31.40	31.08	31.30	31.60
		石油类	0.108	0.101	0.110	0.098	0.097	0.094	0.113
		挥发酚	0.001	0.001	0.001	0.001	0.001	0.001	0.001

表 7.5　天然气开发前、后流域地表水水质监测数据平均值和国标值

单位：mg/L(pH 值除外)

		pH 值	COD_{Mn}	石油类	挥发酚
国家 1 级标准		9.00	15.00	0.050	0.002
国家 2 级标准		9.00	20.00	0.050	0.005
国家 3 级标准		9.00	30.00	0.500	0.010
开发前	绵远河	7.56	14.47	0.231	0.001
	凯　江	9.25	28.89	0.125	0.001
	石亭江	8.08	22.48	0.120	0.001
开发后	绵远河	8.16	10.37	0.071	0.001
	凯　江	7.31	20.07	0.059	0.001
	石亭江	7.41	31.04	0.103	0.001

解　(1)根据主成分分析法，对区域开发前、后的监测数据和国家地表水环境质量标准(GB 3838—2002)中有关指标的标准值进行标准化处理，得到标准矩阵 $\boldsymbol{A}_1$(开发前)、$\boldsymbol{A}_2$(开发后)。

$$\boldsymbol{A}_1=\begin{pmatrix} 0.5253 & -0.1234 & -0.7584 & -0.3689 \\ 0.5253 & -0.2716 & -0.7584 & 0.4611 \\ 0.5253 & 1.2319 & 1.8803 & 1.8443 \\ -1.6258 & -1.1031 & 0.3030 & -0.6455 \\ 0.8988 & 1.0650 & -0.3186 & -0.6455 \\ -0.8490 & 0.1012 & -0.3479 & -0.6455 \end{pmatrix}$$

$$\boldsymbol{A}_2=\begin{pmatrix} 0.8503 & -0.7454 & -0.4990 & -0.3689 \\ 0.8503 & -0.1324 & -0.4990 & 0.4611 \\ 0.8503 & 1.0936 & 2.0286 & 1.8443 \\ -0.1899 & -1.3131 & -0.3810 & -0.6455 \\ -1.2424 & 1.0650 & -0.3186 & -0.6455 \\ -0.8490 & 0.1012 & -0.3479 & -0.6455 \end{pmatrix}$$

(2)分别求上述标准化矩阵的协方差矩阵 $\boldsymbol{C}_1$，$\boldsymbol{C}_2$。

$$C_1=\begin{bmatrix} 0.9999 & 0.5263 & -0.0585 & 0.4069 \\ 0.5263 & 1.0000 & 0.5179 & 0.4967 \\ -0.0585 & 0.5179 & 1.0000 & 0.7265 \\ 0.4069 & 0.4967 & 0.7265 & 1.0000 \end{bmatrix}$$

$$C_2=\begin{bmatrix} 1.0000 & -0.1559 & 0.3462 & 0.6586 \\ -0.1559 & 1.0000 & 0.5933 & 0.4740 \\ 0.3462 & 0.5933 & 1.0000 & 0.8721 \\ 0.6586 & 0.4740 & 0.8721 & 1.0000 \end{bmatrix}$$

（3）分别求其特征矩阵 $\boldsymbol{U}_1$，$\boldsymbol{U}_2$ 和特征值。

$$U_1=\begin{bmatrix} 0.4644 & 0.2569 & 0.7670 & 0.3606 \\ -0.3410 & -0.7428 & 0.2005 & 0.5407 \\ 0.6312 & -0.0849 & -0.5881 & 0.4985 \\ -0.5199 & 0.6124 & -0.1601 & 0.5730 \end{bmatrix}$$

$$U_2=\begin{bmatrix} -0.3909 & 0.4510 & 0.7164 & 0.3614 \\ -0.1576 & 0.6171 & -0.6683 & 0.3845 \\ -0.4691 & -0.6438 & -0.1465 & 0.5865 \\ 0.7761 & -0.0368 & 0.1366 & 0.6145 \end{bmatrix}$$

计算所得的特征值列于表 7.6，可以看到，对于项目开发前的监测结果，3 个特征值所对应的累计贡献率 $\eta_3=100.00\%$，如果选用前两个主成分，那么它们所携带的数据信息已经完全包括了原有所有数据特征的 93.02%，而项目开发后的监测数据处理后，前两个主成分代表了 95.09%的信息。

表 7.6　　特征值、贡献率及累计贡献率一览表

	项目开发前			项目开发后		
序号	λ_1	λ_2	λ_3	λ_1	λ_2	λ_3
特征值	2.0616	0.7289	0.2094	2.2850	0.5676	0.1475
贡献率%	68.72	24.30	6.98	76.17	18.92	4.92
累计贡献率%	68.72	93.02	100.00	76.17	95.09	100.00

根据主成分 F_1，F_2 和对应的均值权数 a_1，a_2 之积 $F_{1\sim2}=\sum_{i=1}^{2}a_iF_i$ 计算，最后得到各水域的综合主成分 $F_{1\sim2}$ 及其排序（表 7.7）。从表中可以看到，在项目进行之前，石亭江的水质最好，优于国家 2 级标准，绵远河水质介于国家 2，3 级

标准之间，而凯江水质介于国家 3，4 级标准之间；项目竣工验收时，区域内所有的水质都已经达到了国家 2 级标准以上，凯江水质最优，接下来是石亭江和绵远河。由此可知，区域开发实施未对区域地表水环境产生负面影响，而且区域开发进行时实施的环境保护设备还有助于区域环境质量的改善。

表 7.7　　各水域综合主成分 $F_{1\sim2}$ 及其排序

		国家 2 级标准	国家 3 级标准	国家 4 级标准	绵远河	凯江	石亭江
开发前	$F_{1\sim2}$	−0.368 6	0.155 7	1.471 9	−1.248 7	0.546 1	−0.556 6
	排序	2	4	6	3	5	1
开发后	$F_{1\sim2}$	0.025 3	0.405 6	1.869 1	−0.538 1	−0.976 4	−0.785 7
	排序	4	5	6	3	1	2

通过四川省德阳地区孝泉—新场—合兴场气田开发工程开发前后区域地表水环境质量的分析，可以看出主成分分析法对于评估区域水环境质量是可行的、有效的，且非常适用，能够在原始信息损失最少的情况下，综合考虑各个参数的结果，通过主成分分析对区域内不同部分的水环境质量进行分级评估，通过排序，可以知道区域内不同区块的波动情况。主成分分析法比原来简单地和国家环境标准进行比较的评价方法更加全面地了解了区域水环境的变化特征，为区域水环境安全的研究提供了评价依据。

【思考题 7】

1. 试述主成分分析的基本思路。
2. 试述主成分的几何意义。
3. 以国家地表水环境质量标准为依据（表 7.8），将水环境质量划分为 5 个等级。具体的水环境数据来自于 2004 年国家水环境质量状况公报，选取三个指标作为评价环境质量的依据。具体数据见表 7.9，试用主成分分析评价各湖区地表水环境质量的好坏。

- 求样本的相关矩阵 $\boldsymbol{R}$；
- 求 $\boldsymbol{R}$ 的特征值及其特征向量；
- 求各主成分的累计贡献率；
- 求第一个主成分，并进行分析。

表 7.8　　地表水环境质量标准

标准等级	高锰酸盐指数	总磷/（mg·L⁻¹）	总氮/（mg·L⁻¹）
Ⅰ类	2	0.010	0.2
Ⅱ类	4	0.025	0.5
Ⅲ类	6	0.050	1.0
Ⅳ类	10	0.100	1.5
Ⅴ类	15	0.200	2.0

表 7.9　　2004年太湖湖体主要污染指标浓度

湖区	高锰酸盐指数	总磷/（mg·L⁻¹）	总氮/（mg·L⁻¹）
五里湖	7.1	0.144	7.00
梅梁湖	5.7	0.102	5.27
西部沿岸区	5.4	0.107	3.33
东部沿岸区	4.0	0.056	1.71
湖心区	4.2	0.059	1.90
全湖平均	4.7	0.078	2.82

4. 某地区22个样地A层土壤重金属含量测定结果见表7.10，试用主成分分析对各样地的7种重金属元素进行评价。

表 7.10　　土壤重金属测定结果　单位：mg/km

序号	Cd	Cr	Cu	Ni	Pb	Hg	As
1	0.221	89.52	42.66	32.63	46.50	0.530	10.90
2	0.462	57.21	46.49	25.42	27.35	0.082	7.82
3	0.132	73.28	31.40	34.38	37.98	0.370	11.47
4	0.109	57.88	25.70	25.82	31.11	0.114	7.54
5	0.078	44.57	36.60	22.06	22.65	0.187	7.39
6	0.129	63.34	22.63	26.85	23.86	0.033	6.90
7	0.132	74.83	18.57	31.71	32.54	0.137	9.08
8	0.170	73.32	56.27	41.84	27.45	0.746	10.46

续表

序号	Cd	Cr	Cu	Ni	Pb	Hg	As
9	0.202	86.26	63.34	51.04	33.42	0.304	10.70
10	0.119	68.62	12.45	25.79	28.23	0.056	7.07
11	0.063	35.39	13.58	16.17	18.29	0.167	12.15
12	0.142	68.41	29.18	33.33	28.96	0.072	10.67
13	0.134	85.39	26.60	37.90	40.04	0.290	6.60
14	0.051	24.61	10.69	13.80	17.65	0.184	8.49
15	0.038	42.23	5.51	10.20	11.24	0.036	5.08
16	0.121	49.73	37.14	32.78	21.41	0.579	5.62
17	0.047	26.93	8.79	10.64	15.71	0.029	10.49
18	0.065	60.17	13.86	18.48	21.04	0.091	10.07
19	0.065	35.84	11.64	17.23	21.37	0.055	8.50
20	0.044	34.19	15.69	12.97	9.80	0.031	9.86
21	0.055	30.19	9.96	13.42	13.03	0.046	10.09
22	0.058	27.78	10.99	15.65	14.19	0.034	13.08

【参考文献】

[1] 何晓群. 现代统计分析方法与应用 [M]. 北京：中国人民大学出版社，2003.

[2] 陈玉成，吕宗清，李章平. 环境数学分析 [M]. 重庆：西南师范大学出版社，1998.

[3] 周晓东，胡振琪. 石油天然气开发对生态环境的破坏与治理 [J]. 资源产业，2000，7：33-36.

[4] 师春元，向启贵. 天然气勘探开发工程环境影响评价 [J]. 石油与天然气化工，2001，30(4)：212-215.

第 8 章　环境因子分析

在环境科学中，我们经常会遇到环境因素众多的数据处理和分析问题。对这些众多的因素，如何提取主要因子，找出每个主因子的明确意义，以便对实际环境问题进行系统分析与评价，这在实际应用中非常重要。因子分析的概念起源于 20 世纪初 K. Pearson，C. Spearmen 等人关于智力测验的统计分析。近年来，随着电子计算机的普及，人们将因子分析的理论广泛应用于环境、资源、气象及经济等领域。本章重点阐述因子分析的基本原理、因子分析模型的构建和求解，并结合环境案例分析该方法在环境科学中的应用。

本章的主要内容是：

- 因子分析概述；
- 正交因子模型；
- 正交因子模型的统计意义；
- 正交因子模型的求解；
- 因子旋转；
- 因子得分；
- 环境应用。

8.1　因子分析概述

因子分析的基本思想是根据相关性大小把变量分组，使得同组内的变量之间相关性较高，但不同组的变量相关性较低。每组内的变量代表一个基本结构，这个基本结构称为公共因子。对于所研究的问题，可试图用最少个数的不可测的所谓公共因子的线性函数与特殊因子之和来描述原来观测的每一分量，即因子分析是将原变量重新进行因子分解，利用数学工具将众多的原变量变换成由少数独立的新变量组成，这种新变量称为因子。因子分析就是找出这些影响系统的最少的独立变量的因子，用较少具有代表性的因子来概括多变量所提供的信息，找出影响观测数据的主要因素，反映环境间内在的关系。

因子分析是从所研究的全部原始变量中将有关信息集中起来，通过探讨相关矩阵的内部依赖结构，将多变量分解成少数因子，以再现原始信息之间的内在关系，并进一步探讨产生这些相关关系的内在原因的一种多元统计分析方法。因子

分析可分解为公共因子和特殊因子两部分，它们客观存在，但又不能直接被测量到。例如在环境统计中，描述环境污染现象的指标有很多，甚至多到几十个。通过因子分析，可以从错综复杂的环境现象中找出少数几个主要因子(方面)，例如对大气粉尘污染的主要污染源的因子分析，从旋转后的因子载荷矩阵可以找出大气颗粒来源主要来自燃油的作用(方面)、燃煤效应(方面)、风沙尘土(方面)三个因素。为此可以帮助我们对复杂现象产生的原因进行分析和解释。

因子分析还可以对变量或样本进行分类。根据因子得分值，在因子轴所构成的空间中把变量或样本点画出来，形象直观地达到分类的目的。

因子分析一般分为两类，一类是研究变量(指标)之间相互关系的 R 型因子分析，另一类则是研究样本之间相互关系的 Q 型因子分析。前者出发点是实测指标间相关系数组成的矩阵，而后者则建立于样本间相似系数组成的矩阵。下面着重介绍 R 型因子分析。

8.2 正交因子模型

因子分析有确定的模型，初学因子分析最大的困难在于理解它的模型，为了理解因子分析的模型，我们从一个例子入手。

例 8.1 城市环境质量评价指标有：COD、BOD_5、NH_3、TSP、SO_2 和 NO_x，现有 100 个样本，用 $\boldsymbol{X}^{(l)}=(x_{l1}, x_{l2}, \cdots, x_{l6})'(l=1, 2, \cdots, 100)$来表示，由 $\boldsymbol{X}^{(l)}(l=1, 2, \cdots, 100)$求得样本的相关矩阵 $\boldsymbol{R}=(r_{ij})_{6\times6}$，其中 r_{ij} 为第 i 个指标与第 j 个指标间的样本相关系数。$\boldsymbol{R}$ 的具体结果如下：

$$\boldsymbol{R}=\begin{pmatrix} 1.00 & & & & & \\ 0.72 & 1.00 & & & & \\ 0.63 & 0.57 & 1.00 & & & \\ 0.09 & 0.16 & 0.14 & 1.00 & & \\ 0.09 & 0.16 & 0.15 & 0.57 & 1.00 & \\ 0.00 & 0.09 & 0.09 & 0.63 & 0.72 & 1.00 \end{pmatrix}$$

从相关矩阵 $\boldsymbol{R}$ 可以看出，前三个指标中两两指标之间的相关系数比较大，后三个指标中两两指标之间的相关系数也比较大，但是前三个指标与后三个指标之间的相关性较小。这说明前三个指标说明了一种原因，后三个指标说明了另一种原因。前者是水环境因素，用 f_1 表示；后者是大气环境因素，用 f_2 表示。若用 x_i 表示第 i 个指标值，则 x_i 可以表示为这两个公共因子的线性组合，即：

$$x_i=\mu_i+a_{i1}f_1+a_{i2}f_2+u_i \tag{8.1}$$

其中，f_1，f_2 称为公共因子，反映了指标所反映的公共因素；a_{i1}，a_{i2} 称为因子载荷，反映了第 i 指标所反映的水环境、大气环境作用的大小；u_i 称为特殊因子，是第 i 指标所反映的特有的原因，这种原因不能被公共因子所反映；μ_i 是第 i 指标值的总平均。记：

$$\boldsymbol{X}=(x_1,\ x_2,\ \cdots,\ x_6)',\ \boldsymbol{A}=(a_{ij})_{6\times 2}$$

$$\boldsymbol{\mu}=(\mu_1,\ \mu_2,\ \cdots,\ \mu_6)',\ \boldsymbol{F}=(f_1,\ f_2)'$$

$$\boldsymbol{U}=(u_1,\ u_2,\ \cdots,\ u_6)'$$

则式(8.1)可以表示为：

$$\boldsymbol{X}=\boldsymbol{\mu}+\boldsymbol{AF}+\boldsymbol{U}$$

称其为因子模型。更一般地，有下述定义：

设样本观察数据 Z，由 p 个变量构成，即 $\boldsymbol{Z}=(z_1,\ z_2,\ \cdots,\ z_p)'$。样本观测数据进行标准化处理，具有均值 0 和方差 1，用 $\boldsymbol{X}=(x_1,\ x_2,\ \cdots,\ x_p)'$ 表示，其协方差矩阵为 $\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$。设标准化后的 p 个变量有 k 个公共因子，用 $\boldsymbol{F}=(f_1,\ f_2,\ \cdots,\ f_k)'$ 表示，其中 $k<p$。标准化后的特殊因子用 $\boldsymbol{U}=(\varepsilon_1,\ \varepsilon_2,\ \cdots,\ \varepsilon_p)'$ 表示。那么标准化后的观察数据可以用公共因子和特殊因子线性表示：

$$\begin{aligned} x_1&=a_{11}f_1+a_{12}f_2+a_{13}f_3+\cdots+a_{1k}f_k+\varepsilon_1\\ x_2&=a_{21}f_1+a_{22}f_2+a_{23}f_3+\cdots+a_{2k}f_k+\varepsilon_2\\ &\cdots\\ x_p&=a_{p1}f_1+a_{p2}f_2+a_{p3}f_3+\cdots+a_{pk}f_k+\varepsilon_p \end{aligned} \tag{8.2}$$

则式(8.2)可以表示为：

$$\boldsymbol{X}=\boldsymbol{AF}+\boldsymbol{U} \tag{8.3}$$

其中，$\boldsymbol{A}=(a_{ij})_{p\times k}$ 为常数矩阵，$\boldsymbol{F}$ 为 k 维向量，可以是随机的也可以是非随机的，$\boldsymbol{U}$ 为 p 维随机向量；称 $\boldsymbol{X}$ 为有 k 个因子的模型，$\boldsymbol{F}$ 为公共因子，$\boldsymbol{U}$ 为特殊因子，$\boldsymbol{A}$ 为因子载荷矩阵。

当 F 是随机向量时，通常假定

$$\begin{aligned} &E(\boldsymbol{F})=\boldsymbol{0},\ \mathrm{Cov}(\boldsymbol{F})=\boldsymbol{I}_k\\ &E(\boldsymbol{U})=\boldsymbol{0},\ \mathrm{Cov}(\boldsymbol{U})=\mathrm{diag}(\psi_1^2,\ \cdots,\ \psi_p^2)\equiv\boldsymbol{\psi}\\ &\mathrm{Cov}(\boldsymbol{F},\ \boldsymbol{U})=\boldsymbol{0} \end{aligned} \tag{8.4}$$

满足式(8.3)，(8.4)的因子模型称为正交因子模型，此时 $\boldsymbol{F}$ 的分量是满足正交条件的。

8.3 正交因子模型的统计意义

从上节的正交因子模型，可以得到：

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}(\boldsymbol{X}) = E(\boldsymbol{XX}') \\ &= E[(\boldsymbol{AF}+\boldsymbol{U})(\boldsymbol{AF}+\boldsymbol{U})'] \\ &= \boldsymbol{A}E(\boldsymbol{FF}')\boldsymbol{A}' + E(\boldsymbol{UU}') = \boldsymbol{AA}' + \boldsymbol{\psi}\end{aligned} \tag{8.5}$$

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{F}) = E(\boldsymbol{XF}') = E[(\boldsymbol{AF}+\boldsymbol{U})\boldsymbol{F}'] = \boldsymbol{A} \tag{8.6}$$

由式(8.5)可以得到：

$$\sigma_{ii} = D(x_i) = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2 + \psi_i^2 \quad (i=1, 2, \cdots, p) \tag{8.7}$$

$$\sigma_{im} = \mathrm{Cov}(x_i, x_m) = a_{i1}a_{m1} + a_{i2}a_{m2} + \cdots + a_{ik}a_{mk} \quad (i \neq k) \tag{8.8}$$

由式(8.6)可以得到：

$$\mathrm{Cov}(x_i, f_j) = a_{ij} \quad (i=1, 2, \cdots, p;\ j=1, 2, \cdots, k) \tag{8.9}$$

从式(8.7)可以看出变量 x_i 的方差是由 k 个公因子和 1 个特殊因子提供的，称 a_{ij} 为第 j 个公因子对变量 x_i 的方差贡献。将 $a_{i1}^2+a_{i2}^2+\cdots+a_{ik}^2$ 记为 h_i^2，它表示了 k 个公因子对变量 x_i 的方差贡献总和，并称 h_i^2 为第 i 个变量的共同度，它刚好是载荷矩阵的第 i 行元素的平方和，而 ψ_i^2 为特殊因子提供的方差，称为特殊度，因此式(8.7)可以改写为：

$$\sigma_{ii} = D(x_i) = h_i^2 + \psi_i^2 \tag{8.10}$$

其中，x_i 已经标准化，所以 $\sigma_{ii}=D(x_i)=1$。当 $h_i^2=1$ 时，$\psi_i^2=0$，这说明 x_i 能被所有公因子的线性组合表示；当 h_i^2 接近 0 时，表明公因子对 x_i 的影响不大，此时 x_i 由特殊因子来描述，由此可以看出 h_i^2 反映了变量对公因子依赖的程度。

另一方面考虑某个指定的公因子 f_j 对各个变量 x_1，x_2，$\cdots$，x_p 的影响用

$$g_j^2 = a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2 \tag{8.11}$$

来描述，它刚好是载荷矩阵第 j 列元素的平方和，称 g_j^2 为公因子 f_j 对所有 p 个变量的方差贡献。显然 g_j^2 越大反映了公因子 f_j 对所有变量的贡献越大，它可作为公因子 f_j 重要性的一个度量。从式(8.9)可以看出，a_{ij} 表示了变量 x_i 与公因子 f_j 的相关系数。如果将因子载荷矩阵 **A** 的所有 g_j^2 都计算出来，使其按大小排序，就可以依次提取最有影响的公共因子。

从上面的讨论，可以归纳出正交因子模型中的载荷矩阵 **A** 具有如下的统计意义：

(1) $h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2$，为第 i 个变量的共同度，它度量了变量 x_i 对 k 个

公因子的依赖程度。

(2)$g_j^2=a_{1j}^2+a_{2j}^2+\cdots+a_{pj}^2$，为第 j 个公因子 f_j 对所有变量的贡献，它是公因子 f_j 重要性的一个度量。

(3)a_{ij}是变量 x_i 与公因子 f_j 的相关系数，它的大小为实际工作中解释公因子的含义提供了一种依据。

8.4 正交因子模型的求解

由因子模型我们知道，用 k 个公共因子和特殊因子来研究相关矩阵的内部依赖结构或者说相关关系的内在原因，从本质上讲是要建立统计模型(8.3)和(8.4)，使其满足方差结构(8.5)，即满足：

$$\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{\psi}$$

如果考虑建立因子模型，第一个要考虑的问题是如何估计载荷矩阵 $\boldsymbol{A}$ 和特殊因子方差 $\psi_i^2(i=1, 2, \cdots, p)$。目前已经提出的方法有许多，如主成分法、极大似然法以及主因子解等方法，在这里我们仅介绍主成分法。

设 $\boldsymbol{\Sigma}$ 的特征根为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$，相应的特征向量组成的矩阵为 $\boldsymbol{P}=(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_p)$，记 $\boldsymbol{D}_\lambda=\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$，此时可以得到：

$$\begin{aligned}\boldsymbol{\Sigma}&=\boldsymbol{P}\boldsymbol{D}_\lambda\boldsymbol{P}'=\lambda_1\boldsymbol{e}_1\boldsymbol{e}_1'+\lambda_2\boldsymbol{e}_2\boldsymbol{e}_2'+\cdots+\lambda_p\boldsymbol{e}_p\boldsymbol{e}_p'\\&=(\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_p}\boldsymbol{e}_p)(\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_p}\boldsymbol{e}_p)'\\&=\boldsymbol{A}\boldsymbol{A}'\end{aligned}\tag{8.12}$$

式(8.12)的分解是公因子个数与变量个数一样，特殊因子的方差为0的因子模型的方差结构形式，即：

$$\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{0}=\boldsymbol{A}\boldsymbol{A}'\tag{8.13}$$

因子分析就是要寻找少数几个公因子来解释变量的相关结构，因此式(8.13)的结构形式在实际应用上是无价值的。类似于主成分分析的思想，如果 $\boldsymbol{\Sigma}$ 的最后 $p-k$ 个特征根很小时，在式(8.12)中将 $\lambda_{k+1}\boldsymbol{e}_{k+1}\boldsymbol{e}_{k+1}'+\lambda_{k+2}\boldsymbol{e}_{k+2}\boldsymbol{e}_{k+2}'+\cdots+\lambda_p\boldsymbol{e}_p\boldsymbol{e}_p'$略去，这样我们就得到：

$$\boldsymbol{\Sigma}\approx(\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_k}\boldsymbol{e}_k)(\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_k}\boldsymbol{e}_k)'=\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'\tag{8.14}$$

这里，$\hat{\boldsymbol{A}}$ 是 $p\times k$ 阶矩阵。式(8.14)近似表明了因子模型，式(8.3)中特殊因子是不重要的，能从 $\boldsymbol{\Sigma}$ 的分解中忽略，如果考虑特殊因子，我们可以用 $\mathrm{diag}(\boldsymbol{\Sigma}-\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}')$来估计 $\boldsymbol{\psi}=\mathrm{diag}(\psi_1^2, \psi_2^2, \cdots, \psi_p^2)$。这里 $\hat{\boldsymbol{A}}$ 由式(8.14)定义。此时近似关系为：

$$\boldsymbol{\Sigma} \approx (\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_k}\boldsymbol{e}_k)(\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_k}\boldsymbol{e}_k)' + \boldsymbol{\psi} \quad (8.15)$$

式中：

$\boldsymbol{\psi}=\mathrm{diag}(\psi_1^2, \psi_2^2, \cdots, \psi_p^2)$，$\psi_i^2=\sigma_{ii}-\sum_{j=1}^{k}a_{ij}^2$，因此载荷矩阵 $\boldsymbol{A}$ 的估计为$\hat{\boldsymbol{A}}=(\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_k}\boldsymbol{e}_k)$。

在实际问题中 $\boldsymbol{\Sigma}$ 是未知的，我们用它的协方差来代替。但是因各个变量的量纲不同，我们一般将数据标准化，利用标准化的数据计算出来的样本协方差就是原数据的相关系数矩阵 $\boldsymbol{R}$，这样我们将上述的方法用于 $\boldsymbol{R}$，可以得到近似于式(8.15)的表示：

$$\begin{aligned}\boldsymbol{R} &\approx \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'+\hat{\boldsymbol{\psi}} \\ &=(\sqrt{\hat{\lambda}_1}\hat{e}_1, \sqrt{\hat{\lambda}_2}\hat{e}_2, \cdots, \sqrt{\hat{\lambda}_k}\hat{e}_k)(\sqrt{\hat{\lambda}_1}\hat{e}_1, \sqrt{\hat{\lambda}_2}\hat{e}_2, \cdots, \sqrt{\hat{\lambda}_k}\hat{e}_k)'+ \\ &\quad \mathrm{diag}(\hat{\psi}_1^2, \hat{\psi}_2^2, \cdots, \hat{\psi}_p^2)\end{aligned} \quad (8.16)$$

其中，$\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant \hat{\lambda}_k$，$\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_k$ 分别是 $\boldsymbol{R}$ 的前 k 个较大的特征根和对应的特征向量。特殊方差 $\hat{\boldsymbol{\psi}}$ 用 $\mathrm{diag}(\boldsymbol{R}-\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}')$来估计。因此，由 $\boldsymbol{R}$ 出发的因子分析模型的载荷矩阵的估计为：

$$\hat{\boldsymbol{A}}=(\sqrt{\hat{\lambda}_1}\hat{e}_1, \sqrt{\hat{\lambda}_2}\hat{e}_2, \cdots, \sqrt{\hat{\lambda}_k}\hat{e}_k) \quad (8.17)$$

特殊因子的方差 ψ_i^2 的估计为：

$$\psi_i^2=1-\sum_{j=1}^{k}\hat{a}_{ij}^2 \quad (i=1, 2, \cdots, p) \quad (8.18)$$

其中，$\hat{a}_{ij}$ 为 $\hat{\boldsymbol{A}}$ 的(i, j)元素。

公因子个数 k 的确定仍按照主成分的思想来定，比较理想的情况是只有少数几个公因子对变量的“贡献”很大，注意到在标准化变量的情况下，所有变量的总方差为 $\boldsymbol{R}$ 的迹，即 $\mathrm{tr}(\boldsymbol{R})=p$，第 j 个公因子 f_j 的“贡献”为：

$$\hat{g}_j^2=\sum_{i=1}^{p}\hat{a}_{ij}^2=(\sqrt{\hat{\lambda}_j}\hat{e}_j')(\sqrt{\hat{\lambda}_j}\hat{e}_j)=\hat{\lambda}_j \quad (8.19)$$

因此寻找一个 k 使得：

$$\left(\sum_{i=1}^{k}\hat{\lambda}_i/p\right)\times 100\% \geqslant 80\%(\text{或 }75\%) \quad (8.20)$$

就确定该 k 为公因子数。

例 8.2 我们仍用第 8.2 节例 8.1 的数据作为例子。首先计算相关矩阵的特征值和特征向量，它的特征值及其累计方差贡献率见表 8.1。

表 8.1　　R 的特征值及其累计方差贡献率

特征值	2.605 4	1.971 1	0.453 3	0.437 5	0.275 6	0.257 1
累计方差贡献率/%	43.42	76.27	83.83	91.12	95.71	100.00

第三个因子的方差贡献率相对于前两个因子作用很小，前两个因子的贡献率已经超过 75%，根据式(8.20)，我们选择前两个因子作为公共因子。

根据式(8.17)得到：

$$\boldsymbol{A}=\begin{pmatrix} 0.6384 & -0.6444 \\ 0.6866 & -0.5475 \\ 0.6510 & -0.5201 \\ 0.6542 & 0.5165 \\ 0.6836 & 0.5508 \\ 0.6383 & 0.6445 \end{pmatrix}$$

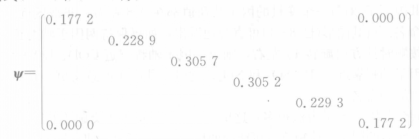
各公共因子对 x_i 的贡献 h_i^2，即第 i 个变量的共同度分别为：

$$\begin{pmatrix} h_1^2 \\ h_2^2 \\ h_3^2 \\ h_4^2 \\ h_5^2 \\ h_6^2 \end{pmatrix}=\begin{pmatrix} 0.8228 \\ 0.7711 \\ 0.6943 \\ 0.6948 \\ 0.7707 \\ 0.8228 \end{pmatrix}$$

第 j 个公因子 f_j 对所有变量的贡献 g_j 分别为：

$$(g_1 \quad g_2)=(2.6054 \quad 1.9711)$$

由式(8.15)和(8.18)得：



$$\boldsymbol{\psi}=\begin{pmatrix} 0.1772 & & & & & 0.0000 \\ & 0.2289 & & & & \\ & & 0.3057 & & & \\ & & & 0.3052 & & \\ & & & & 0.2293 & \\ 0.0000 & & & & & 0.1772 \end{pmatrix}$$

8.5 因子旋转

建立因子分析模型的目的不仅是找出主因子，更重要的是知道每个主因子的明确意义，以便对实际问题进行深入分析。然而用上一节介绍的方法求出的主因子解，若各主因子代表的变量并不很突出，容易使因子的意义含糊不清，不便于对实际问题进行分析。由线性代数知道，一个正交变换对应坐标系的旋转，而且主因子的任意解均可由上述已求得的 $\boldsymbol{A}$ 经过旋转得到，经过旋转后，公共因子对 x_i 的贡献 h_i^2 并不改变，但公共因子本身可能有较大的变化，即 g_j^2 不再与原来的值相同，从而可通过适当的旋转来得到我们比较满意的公共因子。这种变换因子载荷矩阵的方法称为因子旋转(向东进等，2005)。

对于任一正交阵 $\boldsymbol{P}$，由式(8.5)可以得到：

$$\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}' + \boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{P}(\boldsymbol{A}\boldsymbol{P})' + \boldsymbol{\psi} \tag{8.21}$$

另外由式(8.3)可以得到：

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{F}+\boldsymbol{U}=\boldsymbol{A}\boldsymbol{P}(\boldsymbol{P}'\boldsymbol{F})+\boldsymbol{U} \tag{8.22}$$

$$E(\boldsymbol{P}'\boldsymbol{F})=\boldsymbol{P}'E(\boldsymbol{F})=\boldsymbol{0}$$

$$\mathrm{Cov}(\boldsymbol{P}'\boldsymbol{F})=\boldsymbol{P}'\mathrm{Cov}(\boldsymbol{F})\boldsymbol{P}=\boldsymbol{P}'\boldsymbol{I}_k\boldsymbol{P}=\boldsymbol{I}_k$$

$$\mathrm{Cov}(\boldsymbol{P}'\boldsymbol{F},\ \boldsymbol{U})=\boldsymbol{P}'\mathrm{Cov}(\boldsymbol{F},\ \boldsymbol{U})=\boldsymbol{0} \tag{8.23}$$

因此，模型(8.22)仍为正交因子模型，其载荷矩阵为 $\boldsymbol{A}\boldsymbol{P}$，$\boldsymbol{P}'\boldsymbol{F}$ 为新的公共因子，它们是由公共因子 $\boldsymbol{F}$ 旋转得到。新的公共因子的载荷矩阵 $\boldsymbol{A}\boldsymbol{P}$ 与原公共因子的载荷矩阵 $\boldsymbol{A}$ 满足同一个关系式(8.21)，这就说明载荷因子矩阵是不唯一的。这种不唯一性初看起来是不利的，但是正是这种不唯一性可以使我们能够作适当的旋转，使旋转后的公共因子 $\boldsymbol{P}'\boldsymbol{F}$ 能够有更明确的实际意义。

例如，对于上述第 8.4 节例 8.2，按照上节方法所建立的初始载荷矩阵，第一个公共因子 f_1 对每一个变量的因子载荷值都在 0.6 左右，非常接近，不便于解释、命名。由其图形(图 8-1)可直观地看出，如果保持两因子轴的正交关系，将它们按顺时针方向旋转 45°左右，则第一因子轴将靠近 COD、BOD_5 和 NH_3，而第二因子轴则靠近 TSP、SO_2 和 NO_x，这时公共因子的意义更为直观、明确，更易于理解和命名。

旋转公共因子 $\boldsymbol{F}$ 的方法很多，这里只介绍方差最大正交旋转方法。它的基本思想是旋转后的因子载荷矩阵尽可能向两极分化，少数元素取尽可能大的值，而其他元素尽可能接近于零；这样载荷接近于零的因子对该次实验影响不大，而载荷较大的因子应予以重视。

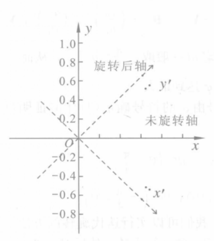


图 8-1 坐标轴旋转图

$$A^* = AP = (a_{ij}^*) \tag{8.24}$$

$$d_{ij} = a_{ij}^* / h_i \quad (j=1, 2, \cdots, k) \tag{8.25}$$

其中，P 为任一 k 阶正交矩阵，用 a_{ij}^* 除以 h_i 是一种类似变量标准化的手法，是为了减小各个变量对公共因子不同程度依赖的影响。所谓方差最大正交旋转法，是选择正交矩阵 P，使得

$$\varphi = \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{p}\left(d_{ij}^2 - \frac{1}{p}\sum_{i=1}^{p}d_{ij}^2\right)^2 \tag{8.26}$$

达到最大(卢崇飞等，1988)。

由于 $d_{ij}^2 \geqslant 0$，因此式(8.26)可能使得某些载荷接近于零，而另外一些载荷较大。当 $k=2$ 时可以准确地求出 P。

P 可以表示成：

$$P = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{8.27}$$

只要求出 θ 就可以确定 P。可以证明，使得 φ 取最大值的旋转角 θ 由下式确定：

$$\tan 4\theta = \frac{D - \frac{2}{p}HG}{C - \frac{1}{p}(H^2 - G^2)} = \frac{\alpha}{\beta} \tag{8.28}$$

其中：

$$D = 2\sum_{i=1}^{p} B_i V_i, \quad H = \sum_{i=1}^{p} B_i, \quad G = \sum_{i=1}^{p} V_i,$$

$$C=\sum_{i=1}^{p}(B_i^2-V_i^2),\ B_i=\left(\frac{a_{i1}}{h_i}\right)^2-\left(\frac{a_{i2}}{h_i}\right)^2,\ V_i=2\frac{a_{i1}}{h_i}\cdot\frac{a_{i2}}{h_i} \quad (8.29)$$

由式(8.28)可以确定 θ(一般取 $-\frac{\pi}{4}\leqslant\theta\leqslant\frac{\pi}{4}$，从而 $-\pi\leqslant 4\theta\leqslant\pi$)，按照此角度作正交旋转，便可以使 φ 达到最大。

可以证明，θ 的符号由 α 的符号确定(具体的证明过程，请参考相关的参考书)。

当 $\alpha>0$，$4\theta\in(0,\pi)$，$\theta\in\left(0,\frac{\pi}{4}\right)$

当 $\alpha<0$，$4\theta\in(-\pi,0)$，$\theta\in\left(-\frac{\pi}{4},0\right)$

对于 $k>2$ 的情况，我们可以实行迭代旋转的方法，即首先对第一和第二两个因子利用上述方法确定 θ 角进行旋转，然后对新的第一因子与原来的第三因子利用同样的方法确定旋转角进行旋转，直至 $\frac{1}{2}k(k-1)$ 对因子都进行旋转完毕，这叫做一个旋转循环。随后又重新开始同样的一个旋转循环，叫做第二循环。这样一轮轮循环重复进行，直至达到某个事先给定的收敛准则为止。

如果我们用 $\varphi_i(i=1,2,\cdots)$ 表示经过 i 轮循环旋转得到的载荷矩阵方差函数 φ 的值，则有 $\varphi_1\leqslant\varphi_2\leqslant\cdots$。

一般地，我们可以在 φ_i 值稳定到一定程度停止运算。把最后得到的 **A** 作为所求的结果。

例 8.3 我们仍用例 8.1 的数据作为例子，将上节得到的载荷矩阵 **A** 进行因子旋转。

未对载荷矩阵 **A** 进行因子旋转之前，方差

$$\varphi=0.006\,0$$

通过式(8.27)、(8.28)和(8.29)，得

$$\theta=0.785\,3$$

即：$\tan 4\theta=-2.578\,6\times10^{-4}$

$$P=\begin{pmatrix}0.707\,2 & -0.707\,1\\ 0.707\,1 & 0.707\,2\end{pmatrix}$$

为此，将上节得到的载荷矩阵 **A** 进行因子旋转后得到：

$$A^{*}=AP=\begin{pmatrix} -0.0042 & -0.9071 \\ 0.0984 & -0.8726 \\ 0.0926 & -0.8281 \\ 0.8278 & -0.0973 \\ 0.8729 & -0.0938 \\ 0.9071 & 0.0045 \end{pmatrix}$$

载荷矩阵方差函数 $\varphi=0.4835$，可见经过因子旋转后，方差变大。

从旋转后的载荷矩阵 A^{*} 的因子载荷可以看出，公共因子 f_1，f_2 反映了指标所反映的公共因素，其中 f_1 反映了 TSP、SO_2 和 NO_x 三个指标所表示的大气环境因素，f_2 反映了 COD、BOD_5 和 NH_3 三个指标所表示的水环境因素。

因子旋转后，各公共因子对 x_i 的贡献 h_i^2，即第 i 个变量的共同度不变，仍为：

$$\begin{pmatrix} h_1^2 \\ h_2^2 \\ h_3^2 \\ h_4^2 \\ h_5^2 \\ h_6^2 \end{pmatrix}=\begin{pmatrix} 0.8228 \\ 0.7711 \\ 0.6943 \\ 0.6948 \\ 0.7707 \\ 0.8228 \end{pmatrix}$$

因子旋转后，第 j 个公因子 f_j 对所有变量的贡献 g_j 分别为：

$$(g_1 \quad g_2)=(2.2883 \quad 2.2882)$$

经过因子旋转特殊向量方差 ψ 也保持不变，仍为：

$$\psi=\begin{pmatrix} 0.1772 & & & & & 0.0000 \\ & 0.2289 & & & & \\ & & 0.3057 & & & \\ & & & 0.3052 & & \\ & & & & 0.2293 & \\ 0.0000 & & & & & 0.1772 \end{pmatrix}$$

8.6 因子得分

到目前为止，我们已经讨论了如何从样本协方差矩阵 $\boldsymbol{\Sigma}$ 或者相关矩阵 $\boldsymbol{R}$ 来获得公共因子和因子载荷，并且知道如何通过因子旋转来确定公共因子的含义。

因子模型建立起来后，我们应当反过来考察每一个样本。例如，分析各区域

污染物成分的因子模型建立之后，我们希望知道每个区域污染状况的轻重，把各区域按污染的轻重划分归类。要解决这个问题，在统计模型上就需要将公共因子用变量的线性组合来表示，也即由原评价指标值来估计它的因子得分。

设公共因子 F 由变量 x 表示的线性组合为：

$$f_j=\beta_{j1}x_1+\beta_{j2}x_2+\cdots+\beta_{jp}x_p \quad (j=1, 2, \cdots, k) \tag{8.30}$$

上式称为因子得分函数，由它来计算每个样本的公共因子得分。如果我们取 2 个公共因子，这样就可以在二维平面上作出因子得分的散点图，进而对样本进行分类或对问题作出更深入的研究。下面我们就讨论因子得分的计算方法。

由于式(8.30)中方程的个数少于变量的个数，因此只能在最小二乘意义下对因子得分进行估计。

据多元回归方程理论可知，用估计量 $\hat{f}_j$ 对式(8.30)中的 f_j 进行估计，且欲求式(8.30)中的 β_{ji}，应有下列正则方程：

$$\begin{cases} r_{11}\beta_{j1}+r_{12}\beta_{j2}+\cdots+r_{1p}\beta_{jp}=l_{1j} \\ r_{21}\beta_{j1}+r_{22}\beta_{j2}+\cdots+r_{2p}\beta_{jp}=l_{2j} \\ \cdots \\ r_{p1}\beta_{j1}+r_{p2}\beta_{j2}+\cdots+r_{pp}\beta_{jp}=l_{pj} \end{cases} \tag{8.31}$$

式中 $l_{ij}(i=1, 2, \cdots, p; j=1, 2, \cdots, k)$ 为变量 x_i 与公因子 f_j 的相关系数，因为各因子相互无关，所以实际上 $l_{ij}=a_{ij}$，即 l_{ij} 是对应因子载荷矩阵 $\boldsymbol{A}$ 中 a_{ij} 元素。而 r_{ij} 是变量间的相关系数矩阵的元素，故式(8.31)的解为：

$$\begin{pmatrix} \beta_{j1} \\ \beta_{j2} \\ \vdots \\ \beta_{jp} \end{pmatrix}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p1} & \cdots & r_{pp} \end{pmatrix}^{-1}\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{pmatrix}$$

故式(8.30)的解为：

$$\hat{\boldsymbol{F}}=\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{X} \tag{8.32}$$

其中，$\boldsymbol{R}$ 为原变量的相关系数矩阵。

$$\hat{\boldsymbol{F}}=(f_1, f_2, \cdots, f_k)'$$
$$\boldsymbol{X}=(x_1, x_2, \cdots, x_p)'$$
$$\boldsymbol{A}'=(a_{ji})_{k\times p}$$

当因子正交时，$\boldsymbol{A}'$为旋转后的因子载荷矩阵 $\boldsymbol{A}$ 的转置。

8.7 环境应用

例 8.4 某地区对城市大气颗粒物进行监测。得到16个样本，样本颗粒中各类物质的含量见表8.2(陈玉成等，1998)。下面我们对该监测数据进行因子分析得到相关结论。

表8.2 样本中大气颗粒物成分分析结果表 单位：mg/kg

序号	Br	K	Ba	Rb	Sc	Fe	Zn	Ni	V	W	As
1	180	11 000	820	58	18.0	22 000	950	110	274	5.9	60
2	97	7 800	650	39	9.6	16 000	930	44	100	6.3	100
3	120	8 600	490	45	8.2	14 000	820	45	107	3.3	72
4	200	7 400	390	31	9.5	13 000	1 500	55	183	10.0	75
5	29	5 400	250	33	5.6	10 000	170	30	88	3.2	25
6	42	9 100	490	43	6.1	14 000	370	17	93	2.5	39
7	60	12 000	520	54	10.0	21 000	780	45	129	4.3	49
8	38	8 700	430	41	8.2	16 000	680	37	96	4.9	56
9	110	5 400	250	30	4.6	7 300	860	39	1	2.7	53
10	38	4 900	174	20	3.5	6 700	480	36	50	3.1	39
11	100	7 100	360	29	5.5	11 000	960	22	28	5.3	25
12	60	4 200	130	15	2.1	4 400	840	17	24	3.9	25
13	15	5 800	240	27	5.5	11 000	650	25	49	4.9	40
14	17	8 000	260	35	5.1	12 000	370	20	48	3.5	30
15	19	870	290	38	5.8	14 000	800	26	40	6.1	25
16	13	46 000	20	20	3.7	7 200	370	14	44	3.7	25

解 1. 对观测数据进行标准化处理，然后把标准化后的数据用矩阵 $\boldsymbol{X}$ 表示。

$$
X=\begin{pmatrix}
1.865 & 0.147 & 2.266 & 1.950 & 2.949 & 1.946 & 0.713 & 3.191 & 2.777 & 0.690 & 0.626 \\
0.443 & -0.170 & 1.428 & 0.348 & 0.710 & 0.720 & 0.651 & 0.331 & 0.225 & 0.903 & 2.429 \\
0.837 & -0.091 & 0.640 & 0.854 & 0.337 & 0.312 & 0.309 & 0.374 & 0.328 & -0.690 & 1.166 \\
2.208 & -0.210 & 0.147 & -0.327 & 0.683 & 0.107 & 2.424 & 0.807 & 1.442 & 2.868 & 1.302 \\
-0.722 & -0.408 & -0.543 & -0.158 & -0.357 & -0.506 & -1.712 & -0.276 & 0.050 & -0.744 & -0.952 \\
-0.499 & -0.041 & 0.640 & 0.685 & -0.223 & 0.312 & -1.090 & -0.840 & 0.123 & -1.115 & -0.321 \\
-0.191 & 0.246 & 0.787 & 1.613 & 0.816 & 1.742 & 0.185 & 0.374 & 0.651 & -0.159 & 0.130 \\
-0.568 & -0.081 & 0.344 & 0.517 & 0.337 & 0.720 & -0.126 & 0.027 & 0.167 & 0.159 & 0.445 \\
0.666 & -0.408 & -0.543 & -0.411 & -0.623 & -1.057 & 0.433 & 0.114 & -1.226 & -1.009 & 0.310 \\
-0.568 & -0.457 & -0.918 & -1.254 & -0.916 & -1.180 & -0.748 & -0.016 & -0.508 & -0.797 & -0.321 \\
0.495 & -0.239 & -0.001 & -0.495 & -0.383 & -0.301 & 0.744 & -0.623 & -0.830 & 0.372 & -0.952 \\
-0.191 & -0.527 & -1.135 & -1.676 & -1.290 & -1.650 & 0.371 & -0.840 & -0.889 & -0.372 & -0.952 \\
-0.962 & -0.368 & -0.593 & -0.664 & -0.383 & -0.301 & -0.220 & -0.493 & -0.522 & 0.159 & -0.276 \\
-0.927 & -0.150 & -0.494 & 0.011 & -0.490 & -0.097 & -1.090 & -0.710 & -0.537 & -0.584 & -0.727 \\
-0.893 & -0.857 & 0.346 & 0.264 & -0.303 & 0.312 & 0.247 & -0.450 & -0.654 & 0.797 & -0.952 \\
-0.996 & 3.614 & -1.677 & -1.254 & -0.863 & -1.078 & -1.090 & -0.970 & -0.596 & -0.478 & -0.952
\end{pmatrix}
$$

2. 求样本的相关矩阵 $\boldsymbol{R}$。

$$
R=\begin{pmatrix}
1.000 & & & & & & & & & & \\
-0.160 & 1.000 & & & & & & & & & \\
0.589 & -0.258 & 1.000 & & & & & & & & \\
0.341 & -0.144 & 0.871 & 1.000 & & & & & & & \\
0.649 & -0.049 & 0.899 & 0.831 & 1.000 & & & & & & \\
0.348 & -0.097 & 0.880 & 0.947 & 0.879 & 1.000 & & & & & \\
0.810 & -0.257 & 0.392 & 0.145 & 0.428 & 0.265 & 1.000 & & & & \\
0.744 & 0.122 & 0.754 & 0.639 & 0.913 & 0.653 & 0.475 & 1.000 & & & \\
0.655 & 0.016 & 0.773 & 0.698 & 0.924 & 0.757 & 0.380 & 0.866 & 1.000 & & \\
0.578 & -0.119 & 0.325 & 0.124 & 0.462 & 0.349 & 0.792 & 0.404 & 0.496 & 1.000 & \\
0.633 & -0.131 & 0.670 & 0.431 & 0.586 & 0.479 & 0.564 & 0.554 & 0.517 & 0.446 & 1.000
\end{pmatrix}
$$

3. 求 $\boldsymbol{R}$ 的特征值 λ 及其相应的特征向量。

$\boldsymbol{R}$ 的特征值及其累计方差贡献率，见表 8.3。

表 8.3 R 的特征值及其累计方差贡献率

特征值	6.535	1.818	1.045	0.576	0.538	0.261	0.116	0.067	0.035	0.005	0.004
累计方差贡献率/%	59.41	75.94	85.43	90.67	95.56	97.94	98.99	99.60	99.92	99.97	100.00

可以看出前三个特征根的累计方差贡献率已超过85%，因此我们选择三个公共因子就可以了。它们对应的特征向量分别为：

$$e_1=(0.303, -0.072, 0.358, 0.307, 0.376, 0.328, 0.241, 0.347, 0.349, 0.227, 0.282)'$$

$$e_2=(-0.349, 0.153, 0.185, 0.403, 0.154, 0.315, -0.540, 0.027, 0.105, -0.451, -0.170)'$$

$$e_3=(0.086, 0.918, -0.180, -0.148, 0.108, -0.058, -0.030, 0.095, 0.225, 0.140, -0.026)'$$

4. 求因子载荷矩阵 **A**。

根据式(8.17)，得：

$$\boldsymbol{A}=(\sqrt{\lambda_1}\boldsymbol{e}_1\ \sqrt{\lambda_2}\boldsymbol{e}_2\ \sqrt{\lambda_3}\boldsymbol{e}_3)=\begin{pmatrix} 0.7746 & -0.4703 & 0.0878 \\ -0.1838 & 0.2068 & 0.9381 \\ 0.9145 & 0.2492 & -0.1844 \\ 0.7847 & 0.5427 & -0.1512 \\ 0.9602 & 0.2076 & 0.1105 \\ 0.8395 & 0.4252 & -0.0596 \\ 0.6157 & -0.7279 & -0.0314 \\ 0.8881 & 0.0365 & 0.0967 \\ 0.8934 & 0.1410 & 0.2299 \\ 0.5802 & -0.6082 & 0.1434 \\ 0.7210 & -0.2290 & -0.0270 \end{pmatrix}$$

共同度，

$$\boldsymbol{h}=\begin{pmatrix} 0.8289 \\ 0.9565 \\ 0.9324 \\ 0.9332 \\ 0.9773 \\ 0.8891 \\ 0.9100 \\ 0.7994 \\ 0.8710 \\ 0.7271 \\ 0.5730 \end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(6.5351 \quad 1.8179 \quad 1.0448)$。

5. 对因子载荷矩阵 **A** 作正交旋转。

运用公式(8.23)～(8.29)对载荷矩阵 **A** 作正交旋转，经过四轮循环后 φ 变化很小，正交循环过程结束。**A** 未进行旋转时 $\varphi_0=0.1825$。具体旋转过程如下：

(1) 第一轮循环过程如下：

①对 **A** 的第 1，2 列进行如下旋转，取：

$$\theta=0.5970$$

则 $\boldsymbol{P}=\begin{pmatrix} 0.8270 & -0.5621 \\ 0.5621 & 0.8270 \end{pmatrix}$

$$\boldsymbol{AP}=\begin{pmatrix} 0.3763 & -0.8244 & 0.0878 \\ -0.0358 & 0.2743 & 0.9381 \\ 0.8964 & -0.3080 & -0.1844 \\ 0.9541 & 0.0077 & -0.1512 \\ 0.9108 & -0.3680 & 0.1105 \\ 0.9333 & -0.1203 & -0.0596 \\ 0.1000 & -0.9481 & -0.0314 \\ 0.7550 & -0.4690 & 0.0967 \\ 0.8182 & -0.3856 & 0.2299 \\ 0.1380 & -0.8291 & 0.1434 \\ 0.4676 & -0.5947 & -0.0270 \end{pmatrix}$$

②对 **A** 的第 1，3 列进行如下旋转，取：

$$\theta=-0.0109$$

则 $\boldsymbol{P}=\begin{pmatrix} 0.9999 & 0.0109 \\ -0.0109 & 0.9999 \end{pmatrix}$

$$AP=\begin{pmatrix} 0.3753 & -0.8244 & 0.0919 \\ -0.0461 & 0.2743 & 0.9376 \\ 0.8984 & -0.3080 & -0.1746 \\ 0.9557 & 0.0077 & -0.1408 \\ 0.9096 & -0.3680 & 0.1205 \\ 0.9339 & -0.1203 & -0.0494 \\ 0.1004 & -0.9481 & -0.0303 \\ 0.7539 & -0.4690 & 0.1049 \\ 0.8156 & -0.3856 & 0.2388 \\ 0.1364 & -0.8291 & 0.1449 \\ 0.4678 & -0.5947 & -0.0219 \end{pmatrix}$$

③对 $\boldsymbol{A}$ 的第 2，3 列进行如下旋转，取：

$$\theta=-0.1556$$

则 $\boldsymbol{P}=\begin{pmatrix} 0.9879 & 0.1549 \\ -0.1549 & 0.9879 \end{pmatrix}$

$$AP=\begin{pmatrix} 0.3753 & -0.8287 & -0.0369 \\ -0.0461 & 0.1257 & 0.9688 \\ 0.8984 & -0.2772 & -0.2202 \\ 0.9557 & 0.0295 & -0.1379 \\ 0.9096 & -0.3822 & 0.0620 \\ 0.9339 & -0.1112 & -0.0674 \\ 0.1004 & -0.9320 & -0.1768 \\ 0.7539 & -0.4796 & 0.0310 \\ 0.8156 & -0.4179 & 0.1762 \\ 0.1364 & -0.8416 & 0.0147 \\ 0.4678 & -0.5841 & -0.1137 \end{pmatrix}$$

经过第一轮循环后，$\boldsymbol{A}$ 经正交旋转得到 $\boldsymbol{A}_1^*$，因子载荷矩阵各列的方差为：

$$\varphi_1=0.3661$$

$$A_1^* = \begin{pmatrix} 0.3753 & -0.8287 & -0.0369 \\ -0.0461 & 0.1257 & 0.9688 \\ 0.8984 & -0.2772 & -0.2202 \\ 0.9557 & 0.0295 & -0.1379 \\ 0.9096 & -0.3822 & 0.0620 \\ 0.9339 & -0.1112 & -0.0674 \\ 0.1004 & -0.9320 & -0.01768 \\ 0.7539 & -0.4796 & 0.0310 \\ 0.8156 & -0.4179 & 0.1762 \\ 0.1364 & -0.8416 & 0.0147 \\ 0.4678 & -0.5841 & -0.1137 \end{pmatrix}$$

共同度不变。各个公因子 f_j 对所有变量的贡献 g=(5.044 0　3.261 5　1.092 3)。

(2)与第一轮循环类似，经过第二轮循环后，A 经正交旋转得到 A_2^*，因子载荷矩阵各列的方差为：

$$\varphi_2 = 0.3663$$

$$A_2^* = \begin{pmatrix} 0.3678 & -0.8321 & -0.0338 \\ -0.0568 & 0.1245 & 0.9684 \\ 0.8983 & -0.2854 & -0.2097 \\ 0.9575 & 0.0206 & -0.1261 \\ 0.9051 & -0.3910 & 0.0725 \\ 0.9336 & -0.1199 & -0.0562 \\ 0.0937 & -0.9326 & -0.1773 \\ 0.7489 & -0.4868 & 0.0394 \\ 0.8094 & -0.4260 & 0.1854 \\ 0.1282 & -0.8429 & 0.0148 \\ 0.4636 & -0.5884 & -0.1091 \end{pmatrix}$$

共同度不变。各个公因子 f_j 对所有变量的贡献 g=(5.009 2　3.301 8　1.086 7)。

(3)经过第三轮循环后，A 经正交旋转得到 A_3^*，因子载荷矩阵各列的方差为：

$$\varphi_3 = 0.3663$$

$$
A_3^* = \begin{pmatrix}
0.3668 & -0.8326 & -0.0338 \\
-0.0569 & 0.1244 & 0.9684 \\
0.8980 & -0.2865 & -0.2094 \\
0.9576 & 0.0194 & -0.1258 \\
0.9045 & -0.3922 & 0.0727 \\
0.9335 & -0.1211 & -0.0559 \\
0.0926 & -0.9327 & -0.1774 \\
0.7482 & -0.4878 & 0.0395 \\
0.8088 & -0.4270 & 0.1856 \\
0.1272 & -0.8430 & 0.0148 \\
0.4629 & -0.5889 & -0.1090
\end{pmatrix}
$$

共同度不变。各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(5.0037\quad 3.3074\quad 1.0866)$。由于第二轮循环与第三轮循环后，因子载荷矩阵各列的方差变化很小，所以此时停止计算。

从旋转后的因子载荷矩阵可以看出该城市的大气颗粒来源主要来自三个因素：第一因素由 Ba、Rb、Sc、Fe、Ni、V 构成，它反映了燃油的作用；第二因素由 Br、Zn、W、As 构成，它反映了燃煤效应；第三因素由 K 构成，它主要是风沙尘土的结果。因此用因子分析方法，可以较好地找出当地大气粉尘污染的主要污染源。

6. 求因子得分。

(1)求特殊向量方差 $\boldsymbol{\psi}$。

$$
\boldsymbol{\psi} = \begin{pmatrix}
0.1711 & & & & & & & & & & 0.0000 \\
& 0.0435 & & & & & & & & & \\
& & 0.0676 & & & & & & & & \\
& & & 0.0668 & & & & & & & \\
& & & & 0.0227 & & & & & & \\
& & & & & 0.1109 & & & & & \\
& & & & & & 0.0900 & & & & \\
& & & & & & & 0.2006 & & & \\
& & & & & & & & 0.1291 & & \\
& & & & & & & & & 0.2729 & \\
0.0000 & & & & & & & & & & 0.4270
\end{pmatrix}
$$

(2) 运用式(8.30)，得到因子得分 $\boldsymbol{F}$，结果见表 8.4。

表 8.4　因子得分表

样本序号	因子得分		
	f_1	f_2	f_3
1	2.518 2	−0.781 9	0.720 5
2	0.568 3	−0.861 5	−0.387 6
3	0.604 7	−0.106 9	−0.262 1
4	−0.396 5	−2.905 7	0.302 9
5	−0.005 7	1.162 1	−0.177 9
6	0.580 7	1.256 5	−0.287 5
7	1.314 3	0.529 0	0.083 2
8	0.507 8	0.288 0	−0.152 0
9	−0.737 8	−0.235 3	−0.642 6
10	−0.829 7	0.366 0	−0.303 4
11	−0.683 6	−0.432 8	−0.401 8
12	−1.577 7	−0.274 3	−0.487 8
13	−0.527 6	0.223 8	−0.341 4
14	−0.109 6	1.040 3	−0.228 6
15	−0.252 4	0.212 7	−0.914 6
16	−0.973 4	0.520 0	3.480 8

例 8.5　为了全面系统地分析评价地表水质量，往往要考虑众多对水质有影响的因素，结合当地的地表水环境质量特点，选取 pH(X_1)、五日生化耗氧量 BOD_5(X_2)、化学耗氧量 COD(X_3)、阴离子洗涤剂(X_4)、非离子氨(X_5)、溶解氧 DO(X_6)、总磷(X_7)、总铅(X_8)、总锌(X_9)、石油类(X_{10})指标作为监测分析项目。得到 6 个样本，监测数据及统计结果详见表 8.5，试用因子分析方法对该地的水质状况进行分析。

表 8.5　　　　某地区地表水水质监测统计结果　单位：mg/L(pH 除外)

样本编号	项目名称									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	7.29	4.01	15.7	0.104	0.019 5	5.62	0.097	0.008 1	0.18	0.05
2	7.16	6.83	19.9	0.219	0.071 9	2.34	0.118	0.008 1	0.08	0.06
3	7.19	6.00	15.5	0.063	0.016 5	5.72	0.103	0.006 8	0.15	0.08
4	7.23	5.83	18.6	0.147	0.055 6	3.89	0.111	0.007 4	0.44	0.06
5	7.19	4.40	16.3	0.066	0.028 5	3.49	0.122	0.008 1	0.24	0.10
6	7.01	9.38	24.3	0.236	0.032 4	1.29	0.154	0.006 8	0.20	0.12

解　1. 对观测数据进行标准化处理，然后把标准化后的数据用矩阵 $\boldsymbol{X}$ 表示。

$$\boldsymbol{X}=\begin{pmatrix} 1.189 & -1.070 & -0.793 & -0.468 & -0.820 & 1.076 & -1.019 & 0.858 & -0.285 & -1.044 \\ -0.195 & 0.391 & 0.448 & 1.063 & 1.580 & -0.787 & 0.025 & 0.858 & -1.101 & -0.675 \\ 0.124 & -0.039 & -0.852 & -1.014 & -0.957 & 1.133 & -0.720 & -1.170 & -0.530 & 0.061 \\ 0.550 & -0.127 & 0.064 & 0.104 & 0.834 & 0.094 & -0.323 & -0.234 & 1.835 & -0.675 \\ 0.124 & -0.868 & -0.616 & -0.974 & -0.408 & -0.133 & 0.224 & 0.858 & 0.204 & 0.798 \\ -1.793 & 1.712 & 1.749 & 1.289 & -0.229 & -1.383 & 1.814 & -1.170 & -0.122 & 1.535 \end{pmatrix}$$

2. 求样本的相关矩阵 $\boldsymbol{R}$。

$$\boldsymbol{R}=\begin{pmatrix} 1.000 & -0.920 & -0.863 & -0.653 & -0.117 & 0.818 & -0.941 & 0.557 & 0.213 & -0.825 \\ -0.920 & 1.000 & 0.916 & 0.799 & 0.278 & -0.753 & 0.816 & -0.651 & -0.145 & 0.574 \\ -0.863 & 0.916 & 1.000 & 0.915 & 0.416 & -0.901 & 0.889 & -0.378 & -0.007 & 0.525 \\ -0.653 & 0.799 & 0.915 & 1.000 & 0.645 & -0.826 & 0.664 & -0.134 & -0.133 & 0.168 \\ -0.117 & 0.278 & 0.416 & 0.645 & 1.000 & -0.552 & 0.158 & 0.299 & 0.095 & -0.302 \\ 0.818 & -0.753 & -0.901 & -0.826 & -0.552 & 1.000 & -0.900 & 0.081 & 0.054 & -0.563 \\ -0.941 & 0.816 & 0.889 & 0.664 & 0.158 & -0.900 & 1.000 & -0.373 & -0.025 & 0.837 \\ 0.557 & -0.651 & -0.378 & -0.134 & 0.299 & 0.081 & -0.373 & 1.000 & -0.136 & -0.500 \\ 0.213 & -0.145 & 0.007 & -0.133 & 0.095 & 0.054 & -0.025 & -0.136 & 1.000 & -0.051 \\ -0.825 & 0.574 & 0.525 & 0.168 & -0.302 & -0.563 & 0.837 & -0.500 & -0.051 & 1.000 \end{pmatrix}$$

3. 求 $\boldsymbol{R}$ 的特征值 λ 及其相应的特征向量。

$\boldsymbol{R}$ 的特征值及其累计方差贡献率，见表 8.6。

表 8.6　特征值及其累计方差贡献率

特征值	5.952 4	1.979 7	1.090 0	0.765 6	0.212 3
累计方差贡献率/%	59.52	79.32	90.22	97.88	100.00

从计算结果可看出，原变量的方差在新变量中的集中度很高，前三个因子其方差的累计贡献率已达到所有因子方差的 90.22%，于是可以用这三个因子来反映原始数据的基本信息。其中第一主因子的贡献率为 59.52%，基本反映了某地区的水质状况和污染情况，而第二和第三主因子的贡献率分别为 19.80% 和 10.90%。因此，可以认为这三个新变量能够完全反映变量的变化所代表的水质状况。

它们对应的特征向量：

$$(\boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \boldsymbol{e}_3)=\begin{pmatrix} -0.394 & -0.163 & -0.116 \\ 0.388 & 0.047 & -0.012 \\ 0.396 & -0.115 & -0.084 \\ 0.335 & -0.352 & 0.026 \\ 0.132 & -0.624 & -0.150 \\ -0.373 & 0.214 & -0.024 \\ 0.389 & 0.093 & -0.005 \\ -0.189 & -0.457 & 0.292 \\ -0.039 & -0.004 & -0.931 \\ 0.283 & 0.425 & 0.054 \end{pmatrix}$$

4. 求因子载荷矩阵 $\boldsymbol{A}$。

根据式(8.17)：

$$\boldsymbol{A}=(\sqrt{\lambda_1}\boldsymbol{e}_1 \ \sqrt{\lambda_2}\boldsymbol{e}_2 \ \sqrt{\lambda_3}\boldsymbol{e}_3)=\begin{pmatrix} -0.961 & -0.230 & -0.122 \\ 0.946 & 0.067 & -0.013 \\ 0.965 & -0.161 & -0.088 \\ 0.817 & -0.496 & 0.027 \\ 0.323 & -0.879 & -0.156 \\ -0.910 & 0.301 & -0.025 \\ 0.949 & 0.131 & -0.006 \\ -0.462 & -0.642 & 0.305 \\ -0.095 & -0.005 & -0.972 \\ 0.692 & 0.599 & 0.057 \end{pmatrix}$$

共同度，
$$h=\begin{pmatrix}0.9914\\0.8989\\0.9651\\0.9147\\0.9006\\0.9200\\0.9181\\0.7192\\0.9544\\0.8397\end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(5.9524\quad 1.9797\quad 1.0900)$。

从因子载荷矩阵可以看出该城市的水质污染来源主要来自三个因素。第一因素由 pH(X_1)、五日生化耗氧量 BOD_5(X_2)、化学耗氧量 COD(X_3)、阴离子洗涤剂(X_4)、溶解氧 DO(X_6)、总磷(X_7)、石油类(X_{10}) 构成，它反映了水体受到有机污染物污染、生活污水污染及上游农业生产污染的作用；第二因素由非离子氨(X_5)、总铅(X_8)构成，它反映了重金属及有机污染物污染的效应；第三因素由总锌(X_9)构成，它主要是无机污染物污染的结果。因此用因子分析方法，可以较好地找出当地水质污染的主要污染源。

5. 求因子得分。

运用式(8.30)，得到因子得分 $\boldsymbol{F}$，结果见表 8.7。

表 8.7　因子得分表

编号	f_1	f_2	f_3
1	−4.305 7	0.629 2	0.597 5
2	−2.681 6	−1.894 1	0.653 7
3	2.100 2	1.520 5	0.661 8
4	−3.349 7	−0.967 5	−2.165 4
5	3.353 8	−0.090 4	0.330 2
6	4.883 0	0.802 3	−0.077 9

【思考题 8】

1. 试述因子分析的基本思想。

2. 比较因子分析和主成分分析模型的关系，说明其异同点。

3. 为较全面评价分析我国 2003～2004 年各地区循环经济发展水平，根据我国地理位置特征和区域经济发展状况，在东部、中部和西部地区各选取 3 个省份作为评价对象，分别为北京、上海、广东、山西、安徽、湖北、重庆、内蒙古和甘肃。同时，为了与全国平均水平作比较，也将“全国平均”作为一个评价对象，具体见表 8.8。试用 R 型因子分析方法对各地区循环经济发展情况进行分类。

(1)求样本的相关矩阵 $\boldsymbol{R}$；

(2)求 $\boldsymbol{R}$ 的特征值及其累计方差贡献率；

(3)求因子载荷矩阵、共同度及各个公因子 f_j 对所有变量的贡献；

(4)求因子得分，并进行分析。

表 8.8　各省份循环经济发展情况

省份	单位 GDP 能耗/(吨标煤·(10^4 元)$^{-1}$)	万元 GDP 用水量/(m^3·(10^4 元)$^{-1}$)	单位面积土地 GDP 产出/(10^4 元·km^{-2})	“三废”综合利用产品产值占工业产值比例/%	环境污染治理投资占 GDP 比例/%
北京	1.29	80.78	2 610.02	0.37	1.53
上海	1.07	158.52	9 042.69	0.22	0.94
广东	0.96	289.79	892.29	0.38	0.70
山西	4.23	183.74	194.14	0.83	1.48
安徽	1.47	435.72	343.45	0.67	0.86
湖北	1.42	384.63	339.45	1.45	0.71
重庆	1.32	253.25	323.98	0.54	1.81
内蒙古	2.43	632.36	23.68	0.61	1.63
甘肃	2.70	781.31	38.58	0.76	1.06
全国平均	1.43	405.32	143.98	0.64	1.40

数据来源：中国统计年鉴、能源统计年鉴和环境统计年鉴。

4. 某化工厂在附近地区挑选有代表性的 8 个大气取样点，测定其中 6 种气体的浓度。具体数据见表 4.14，试用 R 型因子分析方法对当地大气的污染源进行分析。

(1)求样本的相关矩阵 $\boldsymbol{R}$；

(2)求 $\boldsymbol{R}$ 的特征值及其累计方差贡献率；

(3)求因子载荷矩阵、共同度及各个公因子 f_j 对所有变量的贡献；

(4)问经过多少轮旋转后，可以较好地对当地大气的污染源进行分析。

5. 表4.7给出了某地区九个农业区的七项指标，试用 R 型因子分析方法对该地区九个农业区进行分类评价。

(1)求样本的相关矩阵 $\boldsymbol{R}$；

(2)求 $\boldsymbol{R}$ 的特征值及其累计方差贡献率；

(3)解释各公因子所代表的意义；

(4)求因子得分，并进行分类评价。

6. 找一环境问题，建立因子分析模型，并对环境问题进行解释。

【参考文献】

[1] 何晓群. 现代统计分析方法与应用 [M]. 北京：中国人民大学出版社，2003.

[2] 向东进，李宏伟. 实用多元统计分析 [M]. 北京：中国地质大学出版社，2005.

[3] 陈玉成，吕宗清，李章平. 环境数学分析 [M]. 重庆：西南师范大学出版社，1998.

[4] 卢崇飞，高惠璇，叶文虎. 环境数理统计学应用及程序 [M]. 北京：高等教育出版社，1988.

第 9 章　人工神经网络

一元线性回归、多元线性回归等传统统计分析方法，虽然可以解决一些预测问题，但由于它们都要求数据满足正态性、独立性等条件，因而应用起来受到限制。受人类活动、气候、气象等众多因素的影响，环境过程往往是高度复杂的非线性过程，其中存在着大量非线性预测、系统识别、仿真等复杂问题。对于这些复杂问题，如用传统的统计方法则存在着数据不完备，难以选择模型的问题。人工神经网络(artificial neural network，ANN)是一门新兴的学科，从 20 世纪 40 年代提出基本概念以来得到了迅速的发展，以其具有大规模并行处理能力、自适应性能力以及适合于求解非线性、容错性和冗余性等数据处理问题而引起众多领域科学家的广泛关注，现在已经成为计算统计学的一个分支。当传统统计假设条件不满足时，可以采用人工神经网络方法，对数据进行处理和预测。本章重点阐述人工神经网络的基本概念以及常用的几个人工神经网络模型。

本章的主要内容是：

- 人工神经网络概述；
- 人工神经元模型；
- BP 神经网络；
- RBF 神经网络；
- 环境应用。

9.1　人工神经网络概述

人脑神经系统的基本单元是神经细胞，即生物神经元。人脑神经系统大约有 10^{11} 个神经细胞，每个细胞约有 10^4 个通路与其他细胞相连，并且通过突触(一个神经细胞和另一个神经细胞相联系的结构部分)交换信息，整个大脑构成了一个纵横交错的、极其复杂的非线性网络结构。人工神经网络正是在人类对其大脑神经网络认识理解的基础上，人工构造的能够实现某种功能的网络系统。ANN 并不是人脑神经网络系统的真实写照，而是对其作简化、抽象和模拟，是大脑生物结构的数学模型。ANN 由大量功能简单且具有自适应能力的信息处理单元——人工神经元(以下简称为神经元)按照大规模并行方式，通过一定的拓扑结构连接而成。一个人工神经网络的神经元模型和结构描述了一个网络的输入向量转化为

输出向量的过程。这个转化过程从数学的角度来看就是一个计算过程(杨晓华等，2005；丛爽，2003)。

人工神经网络的发展大约经历了半个世纪。一般认为，最早用数学模型对神经系统中的神经元进行理论建模的是美国神经生物学家麦卡洛克(W. McCulloch)和数学家皮茨(W. Pitts)。1943年，他们合作提出了兴奋与抑制型神经元模型，合写了名为 *A Logical Calculus Folders Immanent Nervous Activity* 的开拓性文章，提出了MP模型，首次用简单的数学模型模仿出生物神经元的活动功能。1957年，美国计算机学家罗森布拉特(F. Rosenblatt)提出了著名的感知器(perception)模型。它是一个具有连续可调权值矢量的MP模型，经过训练可以达到对一定的输入矢量模式进行分类和识别的目的。1959年，美国工程师威德罗(B. Widrow)和霍夫(M. Hoff)开发出自适应线性单元(adaline)的网络模型，第一次把神经网络研究从纯理论的研究付诸工程应用，掀起了神经网络研究的第一次高潮。1969年，美国麻省理工学院著名的人工智能专家，人工智能创始人之一明斯基(M. Minsky)和帕伯特(S. Papert)在合著的 *Perception* 书中指出了简单感知器的严重局限性，再加上当时基于语言智能和逻辑数学智能的人工智能很热，导致人工神经网络研究陷于低潮。美国加州理工学院物理学家霍普菲尔德(J. Hopfield)对人工神经网络研究的复苏起到了关键的作用。1982年，他提出了Hopfield网络模型，将能量函数引入到对称反馈网络中，使网络的稳定性有了明确的判据，并利用所建立的网络的神经计算能力来解决条件优化问题。另一个突破性的研究成果是儒默哈特(D. Rumelhart)等人在1986年提出的解决多层神经网络权值修正的算法——误差反向传播算法(error back-propagation algorithm，简称BP算法)，解决了明斯基提出的多层网络的设想问题，使ANN得以全面迅速地恢复发展起来。

人工神经网络的研究和发展经历了几起几伏。它的研究大体上分为4个阶段(闻新等，2003；飞思科技产品研发中心，2003)。表9.1总结了这几个阶段的主要特点。

表 9.1　人工神经网络发展的几个阶段

阶段	时间/年	代表人物/会议	主要贡献
准备阶段	1800	Frued	在精神分析学方面，作了一些初步的工作
	1913	Russell	人工神经网络系统的第一个实践“水力装置”
早期阶段	1943	W. McCulloch 和 W. Pitts	提出了 MP 模型
	1948	Wiener	提出了伺服机反馈自稳定系统的概念
	1949	心理学家 D. O. Hebb	在 *The Organization of Behavior* 书中提出了著名的 Hebb 学习规则
	1957	F. Rosenblatt	提出了感知器(perception)模型
	1959	B. Widrow 和 M. Hoff	开发出自适应线性单元(adaline)的网络模型，第一次把神经网络研究从纯理论的研究付诸工程应用，掀起了神经网络研究的第一次高潮
	1961	Caianiello	发表了神经网络数学的理论著作，研究了细胞有限自动机的理论模型
	1969	M. Minsky 和 S. Papert	在合著的 *Perception* 书中指出了简单感知器的严重局限性，再加上当时基于语言智能和逻辑数学智能的人工智能很热，导致人工神经网络研究陷于低潮
过渡阶段	1972	T. Kohonen	提出了联想记忆理论
	1973～1977	J. Adgerson	把线性联想记忆(LAM)应用到识别、重构和任意可视模式的联想问题上
	1977	Adgerson，Silvetstein	建立了 BSB(brain-state-in-a-box)模型
高潮阶段	1982	美国加州理工学院物理学家 J. Hopfield	提出了离散的神经网络模型，标志着人工神经网络研究高潮的又一次到来
	1984	J. Hopfield	又提出了连续神经网络模型
	1986	M. Rumelhart	提出反向传播学习算法(back-propagation algorithm，简称 BP 算法)
	1987	在美国 Snowbirds 召开了第一次国际神经网络会议	自此以后，各国对神经网络的理论和应用研究迅速发展起来
	1990	IBM 公司	推出了 AS400 工作站，提供了一个自由的神经网络仿真开发环境

ANN 模型有多种形式，它取决于网络拓扑结构、神经元传递函数、学习算法三大要素。ANN 具有以下显著的特点：

(1) 大规模并行计算和分布式存储能力。信息以分布方式存储于整个网络中，即使网络局部受损，也不会对整个网络造成很大影响，还可根据不完整或模糊的信息联想出完整的信息，从而得到正确的输出。ANN 具有并行处理特征，信息处理是在大量单元中并行而有层次地进行的，因此运算速度极快。由于其信息处理能力是由整个网络决定的，所以具有较强的鲁棒性。

(2) 非线性映射能力。ANN 各神经元具有非线性映射特征。虽然 ANN 各神经元的结构和功能简单，但由大量神经元构成的网络系统的行为却是丰富多彩和十分复杂的。ANN 是一个高度复杂的非线性动力系统，具有很强的非线性处理能力。

(3) 自适应、自组织、自学习、联想、容错能力。ANN 可以通过对信息的有监督和无监督学习，调整自身的结构。可以处理各种变化的信息，在处理信息的同时非线性系统本身也在不断变化。ANN 可通过训练样本，根据周围环境来改变自己的网络，可以处理一些环境信息十分复杂，背景知识不清楚，推理规则不明确的问题，如语言、模糊推理、文字识别、医学诊断等。

人工神经网络根据网络结构、状态、学习方式以及系统特点的不同，可分为以下几类：

①结构方式
- 前馈网络(feedforward network，如 BP 网络)
- 反馈网络(feedback network，如 Hopfield 网络)

②状态方式
- 离散型网络(如离散型 Hopfield 网络)
- 连续型网络(如连续型 Hopfield 网络)

③学习方式
- 有监督学习网络(如 BP，RBF 网络)
- 无监督学习网络(如自组织网络)

人工神经网络的详细分类详见有关的参考文献。

本章重点给出 BP 人工神经网络和 RBF 人工神经网络的原理及应用。

9.2 人工神经元模型

人工神经元是人工神经网络的基本单元。在利用人工神经网络解决实际问题之前，首先必须掌握人工神经元的模型。人工神经元的基本结构，见图 9-1。

图 9-1 中，x_1，x_2，…，x_n 是神经元的输入，它可以是来自外界的信息，也可能是另一个神经元的输出；w_1，w_2，…，w_n 是神经元的权值，它表示神经元的连接强度，由神经网络的学习过程决定；θ 是神经元的内部阈值(threshold)；$f(\cdot)$是神经元的激活函数(activation function)(也叫传递函数)，其作用是控制输入对输出的激活作用，把可能的无限域变换到给定的范围输出，对输

入、输出进行函数转换，以模拟生物神经元线性或非线性转移特性。由图 9 - 1 可见，简单神经元主要由权值、阈值和 $f(\cdot)$的形式来定义，它通过对多个输入值与权值乘积和施加线性或非线性函数变换而得到输出值 y：

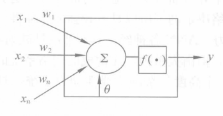


图 9 - 1　人工神经元基本结构图

$$y = f\left(\sum_{i=1}^{n} w_i x_i - \theta\right)$$

$f(\cdot)$一般取下面三种函数：

(1) 线性传递函数（图 9 - 2）

$$y = f(a) = a$$

图 9 - 2　线性传递函数图

(2)双曲正切S型传递函数（图9-3）

$$y=f(a)=\frac{1-\exp(-2a)}{1+\exp(-2a)}$$

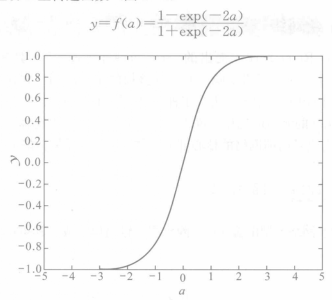


图9-3 双曲正切S型传递函数图

(3)对数S型传递函数（图9-4）

$$y=f(a)=\frac{1}{1+\exp(-a)}$$

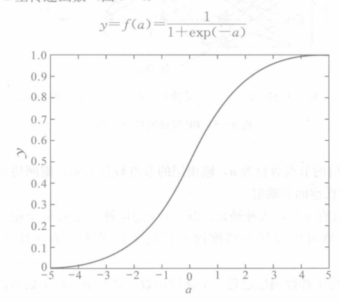


图9-4 对数S型传递函数图

9.3 BP 神经网络

1986 年，D. Rumelhart 等提出的 error back-propagation 算法（简称 BP 算法），系统地解决了多层网络中隐单元层连接权的学习问题，并在数学上给出了完整的推导。目前 BP 模型已成为人工神经网络的重要模型之一，并得到了广泛的应用。在 ANN 的实际应用中，80%～90%的 ANN 模型是采用 BP 网络模型或它的变形，它也是前馈网络的核心部分，体现了 ANN 最精华的部分。

9.3.1 BP 神经网络原理

BP 人工神经网络模型由输入层、隐含层、输出层组成，其拓扑结构如图 9－5 所示。

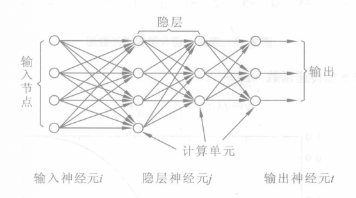


图 9－5　BP 网络的拓扑结构

如果输入层的节点数目为 n，输出层的节点数目为 m，则网络是从 n 维欧氏空间到 m 维欧氏空间的映射。

定理 1　具有 n 个输入神经元、$2n+1$ 个隐层神经元和 m 个输出神经元的前向三层神经网络可以以任意精度逼近任何紧致子集上的连续函数（史忠植，1995）。

定理 1 指出了神经网络逼近一个连续函数，其网络节点个数所需满足的一个充分条件。

定理 2　设 $g(\mathbf{X})$ 为有界单调递增连续函数，I 为 R_n 的紧致子集，固定层数 k

$\geqslant 3$，则对任何连续映射 $f: I \rightarrow R^m$，可由 k 层网络来逼近，此网络的隐单元的输出为 $g(\boldsymbol{X})$，而输入和输出单元之输出关系是线性的。

定理2指出了神经网络逼近一个连续函数，其隐单元、输入和输出单元传递函数所需满足的一个充分条件。

对于一个BP神经网络通常是通过对简单的非线性函数，例如S型函数的复合来实现这一映射的，只要经过少数几次复合，就可得到极复杂的函数，从而可以模拟现实世界的复杂现象。设 $\boldsymbol{X}$ 是 n 维输入向量，$\boldsymbol{Y}$ 是 m 维输出向量，由于对 m 和 n 的大小没有什么限制，使得许多实际环境预测和综合评价问题都可化成用BP神经网络来解决。BP神经网络的这种函数拟合功能，就是它在环境预测和综合评价中应用的理论依据。

9.3.2 BP算法

BP算法的核心是通过一边向后传播误差，一边修正误差的方法来不断调节网络参数(权值和阈值)，以实现或逼近所希望的输入、输出映射关系。它对每一个训练过程进行两次传播计算：第一次，前向计算。从输入层开始向后逐层计算输出，产生最终输出，并计算实际输出与目标输出间的误差；第二次，反向计算。从输出层开始向前逐层传播误差信号，修正权值，直到误差小于给定值。

图9-6给出了BP算法原理图。在这种网络中，学习过程由正向传播和反向传播组成。在正向传播过程中，输入信号从输入层经隐层单元逐层处理，并传向输出层，每一层神经元的状态只影响下一层神经元的状态。如果在输出层不能得到期望的输出，则转入误差反向传播，将输出信号的误差沿原来的连接通路返回。通过修改各层神经元的权值和阈值，使得网络全局误差信号最小。

下面进一步以图9-7所示的三层BP神经网络为例(金菊良等，2000)，详细说明单样本点的BP算法的实现过程。设输入神经元为 h，隐层神经元为 i，输出神经元为 j，n_h，n_i，n_j 分别为三层的节点数目，θ_i，θ_j 分别为隐层节点 i、输出层节点 j 的阈值，w_{hi}，w_{ij} 分别为输入层节点 h 与隐层节点 i 间、隐层节点 i 与输出层节点 j 间的连线的权值，各节点的输入为 x，输出为 y。

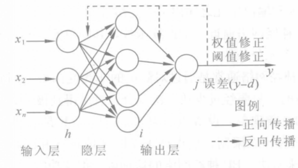


图 9-6 BP 算法原理示意图

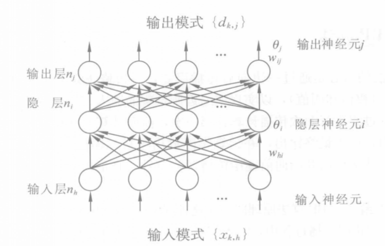


图 9-7 BP 神经网络的拓扑结构

BP 人工神经网络的学习算法包含以下 9 个步骤。

步骤 1： 初始化。为了加快网络的学习效率，需对原始的输入、输出样本作规范化处理。设已归一化的输入、输出样本为 $\{x_{k,h},\ d_{k,j} \mid k=1,\ 2,\ \cdots,\ n_k;\ h=1,\ 2,\ \cdots,\ n_h;\ j=1,\ 2,\ \cdots,\ n_j\}$，$n_k$ 为样本容量。给各连接权 $\{w_{hi}\}$，$\{w_{ij}\}$ 和阈值 $\{\theta_i\}$，$\{\theta_j\}$ 赋予(−0.1，0.1)区间上的随机值。

步骤 2： 置 $k=1$，把样本对$(x_{k,h},\ d_{k,j})$提供给网络$(h=1,\ 2,\ \cdots,\ n_h;\ j=1,\ 2,\ \cdots,\ n_j)$。

步骤 3： 计算隐层各节点的输入 x_i、输出 $y_i(i=1,\ 2,\ \cdots,\ n_i)$。

$$x_i = \sum_{h=1}^{n_h} w_{hi} x_{k,h} + \theta_i$$

$$y_i = f_1(x_i)$$

其中，$f_1(\cdot)$是隐层各节点神经元的激活函数。

步骤 4：计算输出层各节点的输入 x_j、输出 $y_j (j=1, 2, \cdots, n_j)$：

$$x_j = \sum_{i=1}^{n_i} w_{ij} y_i + \theta_j$$

$$y_j = f_2(x_j)$$

其中，$f_2(\cdot)$是输出层各节点神经元的激活函数。

计算第 k 个单样本点的误差：

$$E_k = \sum_{j=1}^{n_j} (y_j - d_{k,j})^2 / 2$$

BP 算法中 E_k 与其他变量之间的函数关系参见图 9-8。

$$x_{k,h} \longrightarrow x_i \xrightarrow{f_1} y_i \longrightarrow x_j \xrightarrow{f_2} y_j \xrightarrow{d_{k,j}} E_k$$

（w_{hi}、θ_i 指向 x_i；w_{ij}、θ_j 指向 x_j）

图 9-8　BP 算法中 E_k 与其他变量之间的函数关系示意图

各层连接权及阈值的调整，按梯度下降法的原则进行。

步骤 5：计算输出层权值和阈值的修正量 Δw_{ij}，$\Delta\theta_j$。

$$\begin{aligned}\text{grad}_{w_{ij}}(E_k) &= \frac{\partial E_k}{\partial w_{ij}} = \frac{\partial E_k}{\partial x_j} \cdot \frac{\partial x_j}{\partial w_{ij}} \\ &= [(y_j - d_{k,j}) \cdot f'_2(x_j)] \cdot y_i \\ &= \delta_{k,j} \cdot y_i\end{aligned}$$

$$\Delta w_{ij} = -\eta \cdot \frac{\partial E_k}{\partial w_{ij}}$$

$$\begin{aligned}\text{grad}_{\theta_j}(E_k) &= \frac{\partial E_k}{\partial \theta_j} = \frac{\partial E_k}{\partial x_j} \cdot \frac{\partial x_j}{\partial \theta_j} \\ &= [(y_j - d_{k,j}) \cdot f'_2(x_j)] \cdot 1 \\ &= \delta_{k,j}\end{aligned}$$

$$\Delta\theta_j = -\eta \cdot \frac{\partial E_k}{\partial \theta_j}$$

步骤 6：计算隐层权值和阈值的修正量 Δw_{hi}，$\Delta\theta_i$。

$$\text{grad}_{w_{hi}}(E_k)=\frac{\partial E_k}{\partial w_{hi}}=\frac{\partial E_k}{\partial x_i}\cdot\frac{\partial x_i}{\partial w_{hi}}$$

$$=\left(\sum_{j=1}^{n_j}\frac{\partial E_k}{\partial x_j}\cdot\frac{\partial x_j}{\partial y_i}\cdot\frac{\partial y_i}{\partial x_i}\right)\cdot\frac{\partial x_i}{\partial w_{hi}}$$

$$=\left[\sum_{j=1}^{n_j}(y_j-d_{k,j})\cdot f'_2(x_j)\cdot w_{ij}\cdot f'_1(x_i)\right]\cdot x_{k,h}$$

$$=\left[\sum_{j=1}^{n_j}\delta_{k,j}\cdot w_{ij}\cdot f'_1(x_i)\right]\cdot x_{k,h}$$

$$\overset{\text{记为}}{=}\delta_{k,i}\cdot x_{k,h}$$

$$\Delta w_{hi}=-\eta\cdot\frac{\partial E_k}{\partial w_{hi}}$$

$$\text{grad}_{\theta_i}(E_k)=\frac{\partial E_k}{\partial \theta_i}=\frac{\partial E_k}{\partial x_i}\cdot\frac{\partial x_i}{\partial \theta_i}$$

$$=\left(\sum_{j=1}^{n_j}\frac{\partial E_k}{\partial x_j}\cdot\frac{\partial x_j}{\partial y_i}\cdot\frac{\partial y_i}{\partial x_i}\right)\cdot\frac{\partial x_i}{\partial \theta_i}$$

$$=\sum_{j=1}^{n_j}(y_j-d_{k,j})\cdot f'_2(x_j)\cdot w_{ij}\cdot f'_1(x_i)$$

$$=\sum_{j=1}^{n_j}\delta_{k,j}\cdot w_{ij}\cdot f'_1(x_i)$$

$$\overset{\text{记为}}{=}\delta_{k,i}$$

$$\Delta\theta_i=-\eta\cdot\frac{\partial E_k}{\partial \theta_i}$$

步骤 7：修正各连接的权值和阈值。

$$w_{ij}^{t+1}=w_{ij}^{t}+\Delta w_{ij}$$
$$\theta_j^{t+1}=\theta_j^{t}+\Delta\theta_j$$
$$w_{hi}^{t+1}=w_{hi}^{t}+\Delta w_{hi}$$
$$\theta_i^{t+1}=\theta_i^{t}+\Delta\theta_i$$

式中，t 为修正次数，学习速率 $\eta\in(0,1)$。若 η 较大，则算法收敛快，但不稳定，可能出现振荡，若 η 较小则算法收敛缓慢。

步骤 8：置 $k=k+1$，取学习模式对$(x_{k,h}, d_{k,j})$提供给网络，转步骤 3，直至全部 n_k 个模式对训练完毕，转步骤 9。

步骤 9：重复步骤 2～8，直至网络全局误差函数。

$$E=\sum_{k=1}^{n_k}E_k=\sum_{k=1}^{n_k}\sum_{j=1}^{n_j}(y_j-d_{k,j})^2/2$$

小于预先设定的一个较小值或学习次数大于预先设定的值，结束学习。

BP 算法程序框图，见图 9-9。

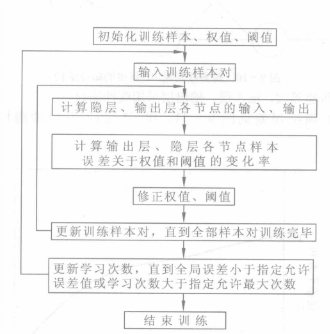


图 9-9　BP 算法程序框图

可见 BP 算法把一组样本的输入、输出问题归纳为一非线性优化问题，它使用了最优化方法中最常用的负梯度下降算法。

用迭代运算求解网络权重和阈值对应于网络的学习记忆过程，加入隐层节点使得优化问题的可调参数增加，从而可得到更精确的解。

BP 算法的优点是算法推导清楚，学习精度较高，可用作一个通用的函数模拟器；从理论上说，用 BP 算法可以逼近任何的非线性函数；经过训练后的 BP 网络运行速度极快，可用于实时处理。但是，BP 算法也可能存在局部极小和收敛缓慢的缺陷。

例 9.1　用 BP 算法解异或问题

异或问题的输入、输出样本对为 (0，0)→0，(0，1)→1，(1，0)→1，(1，1)→0。

解　取隐层节点数目为 2，学习速率为 0.1，BP 网络的拓扑结构见图 9-10。

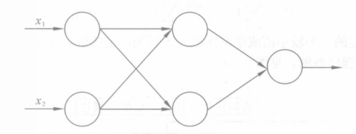


图 9-10 解异或问题 BP 网络的拓扑结构

用 S 型非线性传递函数，输入层、输出层采用线性函数。

BP 网络的训练误差见图 9-11，BP 算法(2-2-1)的隐层训练结果见图9-12。

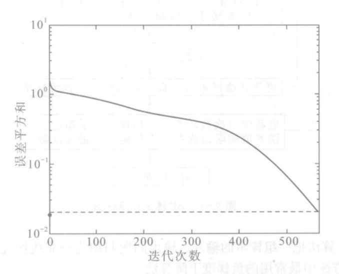


图 9-11 BP 算法(2-2-1)的训练误差图

BP 算法各层训练结果如下：

(1)隐层训练结果

$$w_{hi} = \begin{matrix} 7.7437 & -6.6925 \\ 2.7595 & -2.0861 \end{matrix}$$

$$\theta_i = -1.6911,\ 7.4525$$

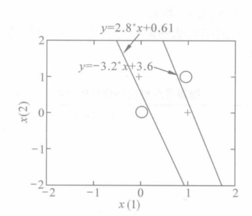


图 9-12 BP 算法(2-2-1)的隐层训练结果图

(2)输出层训练结果

$$w_{ij}=1.6791$$
$$1.7172$$
$$\theta_j=-1.9505$$

BP 计算值：

$$y=0.0271,\ 1.0085,\ 0.8947,\ 0.0886$$

目标值：

$$T=0,\ 1,\ 1,\ 0$$

BP 网络训练结果，见图 9-13。

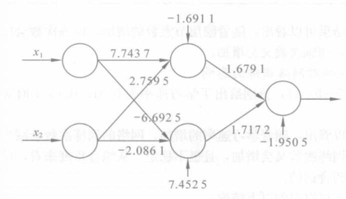


图 9-13 BP 算法(2-2-1)的训练结果图

下面讨论解异或问题的隐层节点数和学习速率对网络训练的影响。

1. 隐层节点对网络训练的影响

BP算法隐层节点对网络训练的影响详见表9.2。BP算法的迭代次数与隐层节点的关系见图9-14。

表9.2 隐层节点对网络训练的影响

隐层节点数	迭代次数	误差
2	568	0.019 7
3	85	0.019 5
4	275	0.019 1
5	141	0.019 7
6	181	0.020 0
7	17	0.019 6
8	42	0.019 6
9	37	0.019 2
10	14	0.016 8
15	55	0.016 5
20	30	0.020 0
25	28	0.020 0
30	59	0.020 0

从上面的结果可以看出，随着隐层节点数的增加，训练次数会减少，但增加到一定数目后，训练次数又会增加。

2. 学习速率对网络训练的影响

如图9-15～9-17，分别给出了学习速率为0.01，0.1，1时对网络训练的影响。

从结果可以看出，随着学习速率的增加，网络的训练次数会减少，但增加到一定数目后，训练次数又会增加，且极不稳定。从综合角度来看，此例学习速率选择0.1是比较合适的。

综上所述，可以得到以下结论：

(1) 太大的学习速率导致学习的不稳定，太小值又导致极长的训练时间。

(2) 在误差一定的情况下，随着隐层节点数的增加，训练次数会减少，但增加到一定数目后，训练次数又会增加。

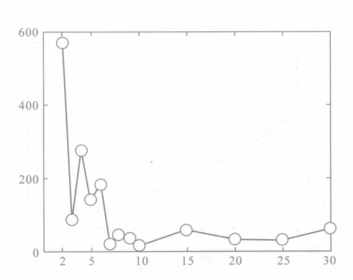


图9-14　BP算法的迭代次数与隐层节点的关系图

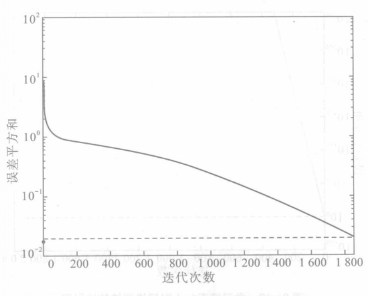


图9-15　学习速率为0.01时网络训练的结果图

（3）BP网络的结构不完全受所要解决问题的限制。网络的输入神经元数目以及输出神经元数目是由问题的要求所决定的，而隐层数是由设计者来决定。

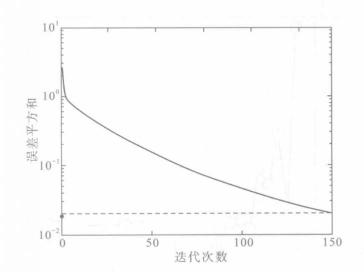


图 9－16　学习速率为 0.1 时网络训练的结果图

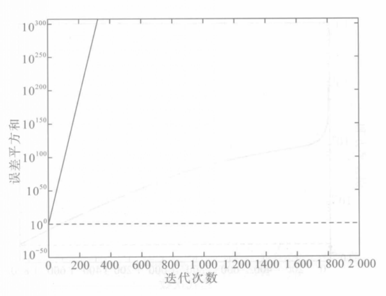


图 9－17　学习速率为 1 时网络训练的结果图

（4）BP 网络的学习采用梯度下降法，网络误差有可能陷入局部极小值，可采用附加动量法等改进方法来训练网络。

9.3.3 环境应用

例 9.2　新疆伊犁河雅马渡站年径流的预测(陈守煜，1997)。该站 23 年实测年径流 y^k 与其相应的 4 个预测因子数据见表 9.3。预测因子 a_1，a_2，a_3，a_4 分别为前一年 11 月至当年 3 月伊犁气象站的总降雨量(mm)，前一年 8 月欧亚地区月平均纬向环流指数，前一年 5 月欧亚地区径向环流指数，前一年 6 月 2 800 MHz 的太阳射电流量(10^{-22} W/m^{2}Hz)。取最前 17 个资料为建模样本，最后 6 年资料为测试样本。BP 神经网络模型的拓扑结构取(4，5，1)。用上述模型运行 66 次，网络的全局误差 E=3.760 57e-008，见表 9.3。表 9.3 同时给出了测试结果和预测值的绝对误差。由此可见所建立的 BP 神经网络模型可以用来预测年径流。

表 9.3　　BP 模型年径流预测的训练和测试结果

	预测因子				年径流值		年径流值误差
	a_1	a_2	a_3	a_4	实测	计算	绝对误差
训练	114.6	1.10	0.71	85.0	346	345.993 9	0.006 1
	132.4	0.97	0.54	73.0	410	409.989 6	0.010 4
	103.5	0.96	0.66	67.0	385	385.008 6	−0.008 6
	179.3	0.88	0.59	89.0	446	445.957 7	0.042 3
	92.7	1.15	0.44	154.0	300	300.021 2	−0.021 2
	115.0	0.74	0.65	252.0	453	453.031 6	−0.031 6
	163.6	0.85	0.58	220.0	495	495.204 5	−0.204 5
	139.5	0.70	0.59	217.0	478	478.235 9	−0.235 9
	76.7	0.95	0.51	162.0	341	341.171 2	−0.171 2
	42.1	1.08	0.47	110.0	326	326.023 9	−0.023 9
	77.8	1.19	0.57	91.0	364	363.983 7	0.016 3
	100.6	0.82	0.59	83.0	456	455.989 8	0.010 2
	55.3	0.96	0.40	69.0	300	299.996 5	0.003 5
	152.1	1.04	0.49	77.0	433	432.984 4	0.015 6
	81.0	1.08	0.54	96.0	336	336.005 2	−0.005 2
	29.8	0.83	0.49	120.0	289	289.131 8	−0.131 8
	248.6	0.79	0.50	147.0	483	482.907 9	0.092 1
测试	89.9	0.96	0.39	105.0	314	317.669 4	3.669 4
	90.0	0.95	0.43	89.0	301	306.747 9	5.747 9

例 9.3 以长江重庆干流段 1989 年的实测水质资料(郭劲松等，2001)来建立 BOD-DO 耦合 BP 网络模型。研究范围为长江重庆干流段，从江津市羊石乡史坝沱到长寿县黄草峡，全长 240.8 km。根据沿江污染源分布状态、水质监测断面以及流场变化情况，将干流段划分为五个研究江段。所采用的实测水质资料如表 9.4 所示，其中 1～4 为枯水期各个河段的水质资料，5～8 为丰水期各个河段的水质资料，两个时期所对应河段分别为：羊石—白沙沱，白沙沱—望龙门，望龙门—寸滩，寸滩—鱼嘴，9 和 10 分别为枯水期与丰水期的鱼嘴—长寿段的水质资料。前四个河段的水质数据作为 BP 网络的训练数据，最后一个河段数据作为验证数据。

表 9.4　BP 网络水质模拟输入水质资料

序号	背景值			河段基本情况				本段污染物负荷值			
	流量 Q/($10^4 m^3 \cdot s^{-1}$)	DO/($mg \cdot L^{-1}$)	BOD/($mg \cdot L^{-1}$)	长度 L/($10^2 km$)	断面宽度 B/km	流速 v/($m \cdot s^{-1}$)	流量 q/($10^2 m^3 \cdot s^{-1}$)	DO/($10^3 kg \cdot (d \cdot km)^{-1}$)	BOD/($10^3 kg \cdot (d \cdot km)^{-1}$)	DO/($mg \cdot L^{-1}$)	BOD/($mg \cdot L^{-1}$)
1	0.299 0	8.6	1.1	1.178	0.55	1.220	0.797 9	0.412 7	0.149 3	8.4	1.1
2	0.306 9	8.4	1.1	0.420	0.30	0.797	0.049 2	0.045 4	1.091 3	8.5	1.2
3	0.307 5	8.5	1.2	0.080	0.35	1.916	6.758 9	67.012 5	14.565 5	8.3	1.5
4	0.442 6	8.3	1.5	0.230	0.35	1.420	0.015 8	0.038 0	0.071 3	8.4	1.2
5	1.720 0	7.5	0.3	1.178	0.80	1.939	2.744 3	0.177 7	0.051 3	7.0	0.6
6	1.747 4	7.0	0.6	0.420	0.70	1.488	0.147 1	0.019 1	0.131 2	7.4	0.8
7	1.748 9	7.4	0.8	0.080	0.70	2.339	0.393 9	2.587 5	0.385 9	7.5	0.9
8	2.146 9	7.5	0.9	0.230	0.70	2.878	0.054 1	0.019 5	0.007 3	7.3	1.0
9	0.442 8	8.4	1.2	0.500	0.30	0.940	0.542 9	0.486 7	0.455 0	8.5	1.5
10	2.147 3	7.3	1.0	0.500	0.80	2.016	1.898 9	0.298 5	0.083 4	7.2	0.9

合理确定 BP 网络的结构是预测性能的基础。经过实验，输入层的神经元数取 9，输出的神经元数取 2，隐含层的神经元数取 10。经过 2 000 次训练后，网络的训练误差为 7.537 39e-008 左右。其训练样本模拟结果和检测样本预测结果与实测值比较，如表 9.5 所示。检测样本模拟结果和检测样本预测结果与实测值

比较，如表9.6所示，可见BP网络可以预测河流水质。

表9.5 训练样本模拟结果与实测值的比较

指标		数据组 1	2	3	4	5	6	7	8
DO	实测值	8.4	8.5	8.3	8.4	7.0	7.4	7.5	7.3
	预测值	8.393 5	8.498 5	8.300 6	8.398 2	7.001 7	7.398 3	7.501 4	7.300 4
	绝对误差	0.006 5	0.001 5	−0.000 6	0.001 8	−0.001 7	0.001 7	−0.001 4	−0.000 4
BOD	实测值	1.1	1.2	1.5	1.2	0.6	0.8	0.9	1.0
	预测值	1.100 1	1.200 0	1.499 7	1.199 3	0.600 2	0.800 3	0.900 4	0.999 7
	绝对误差	−0.000 1	0.000 0	0.000 3	0.000 7	−0.000 2	−0.000 3	−0.000 4	0.000 3

表9.6 检测样本预测结果与实测值的比较

指标	实测值	预测值	绝对误差
DO	8.5	8.307 8	−0.192 2
	7.2	7.015 3	−0.184 7
BOD	1.5	1.222 3	−0.277 7
	0.9	1.017 9	0.117 9

9.4 RBF神经网络

BP神经网络用于预测、评价、函数逼近时，权值调节采用的是负梯度下降法，该法有它的局限性，即存在收敛速度慢和局部极小等缺点。而径向基函数(radial basis function，RBF)神经网络无论在逼近能力、分类能力和学习速度等方面均优于BP神经网络(闻新等，2003)，本节给出Matlab6.5环境下的RBF网络模型。

9.4.1 RBF神经网络原理

RBF网络由两层组成，第一层为隐含的径向基层，第二层为输出线性层，其网络结构如图9-18所示。

径向基函数是径向对称的，最常用的是高斯函数：

$$R_i(\boldsymbol{x})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_i\|^2}{2\sigma_i^2}\right) \qquad (i=1,\ 2,\ \cdots,\ p)$$

其中，$\boldsymbol{x}$ 是 m 维输入向量，$\boldsymbol{c}_i$ 是第 i 个基函数的中心，σ_i 是第 i 个感知的变量，p 是感知单元的个数，$\|\boldsymbol{x}-\boldsymbol{c}_i\|$ 是向量 $\boldsymbol{x}-\boldsymbol{c}_i$ 的范数。

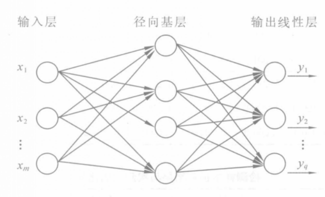


图 9-18　RBF 网络结构

从图 9-18 可以看出，输入层实现从 $\boldsymbol{x}\rightarrow R_i(\boldsymbol{x})$ 的非线性映射，输出层实现从 $R_i(\boldsymbol{x})\rightarrow y_k$ 的线性映射，即：

$$y_k=\sum_{i=1}^{p} w_{ik}R_i(\boldsymbol{x}) \qquad (k=1,\ 2,\ \cdots,\ q)$$

其中，q 是输出节点数。

从理论上而言，RBF 网络可以逼近任何的非线性函数。

9.4.2　RBF 神经网络模型

RBF 网络模型不仅可以用来函数逼近，还可以进行预测和评价。为了能够具体说明 RBF 网络模型的建立过程，本节直接给出用于时间序列预测的 RBF 网络模型。

实际时间序列在时序上常常表现出弱相依性、突变性和随机性等复杂非线性特征，至今对此进行有效描述的数学模型仍不很成熟。TAR 模型、多元线性回归、灰色模型、投影寻踪回归、未确知模拟模型、模糊预测、人工神经网络、组合预测、混沌分析等预测模型(金菊良等，1999)都有各自的特点。

本节将 RBF 网络法应用于时间序列预测，用自相关分析技术分析时间序列

的延迟特性，据此确定 RBF 网络的输入、输出向量，建立了 Matlab6.5 环境下的 RBF 网络模型，并用实例进行了验证。RBF 网络模型包含以下 3 步。

1. 用自相关分析技术确定 RBF 网络模型的输入、输出向量

设时序$\{x^*(i)\}$延迟 k 步的自相关系数$R(k)$为：

$$R(k)=\sum_{i=k+1}^{n}[x^*(i)-e_x][x^*(i-k)-e_x]/\sum_{i=1}^{n}[x^*(i)-e_x]^2$$

$$e_x=\sum_{i=1}^{n}x^*(i)/n$$

其中，n 为实测时序$\{x^*(i)\}$的容量，$k=1,2,\cdots,n_k<[n/10]$ 或 $[n/4]$。$R(k)$的方差随 k 的增大而增大，$R(n_k)$的估计精度随 n_k 的增加而降低，因此 n_k 应取较小的数值。

根据$R(k)$的抽样分布理论，在置信水平 $1-\alpha$ 的情况下，当自相关系数值

$$R(k)\notin[(-1-u_{\alpha/2}\cdot(n-k-1)^{0.5})/(n-k),\ (-1+u_{\alpha/2}\cdot(n-k-1)^{0.5}/(n-k)] \tag{9.1}$$

时，则推断时序$\{x^*(i)\}$延迟 k 步相依性显著，否则时序$\{x^*(i)\}$延迟 k 步相依性不显著。其中，分位数 $u_{\alpha/2}$ 可从正态分布表中查得。它的自回归系数项应与这些相依性显著的延迟步数相对应。设最大相依性延迟步数为 m，则对于 n 个容量的时间序列，其 RBF 网络训练样本的输入、输出向量为以下 $n-m$ 组：

$$\bar{\boldsymbol{x}}=[x^{m+1},x^{m+2},\cdots,x^n],\ \bar{\boldsymbol{y}}=[y^{m+1},y^{m+2},\cdots,y^n] \tag{9.2}$$

其中，$\boldsymbol{x}^i=[x^*(i-m),x^*(i-m+1),\cdots,x^*(i-1)]'$，$\boldsymbol{y}^i=x^*(i)\quad(i=m+1,m+2,\cdots,n)$，$\boldsymbol{x}^i$，$\boldsymbol{y}^i$ 分别为 m 维输入向量和 1 维输出向量，本节输出节点数 $q=1$。$\bar{\boldsymbol{x}}$，$\bar{\boldsymbol{y}}$ 分别为 $n-m$ 组 m 维输入向量和 1 维输出向量所构成的训练样本矩阵。

2. 用 newrb 函数设计一个满足一定精度要求的 RBF 网络

格式：net=newrb($\bar{\boldsymbol{x}}$，$\bar{\boldsymbol{y}}$，*goal*，*spread*)

用 RBF 网络逼近函数时，newrb 可自动增加 RBF 网络的隐层神经元，直到均方误差满足为止。其中 $\bar{\boldsymbol{x}}$，$\bar{\boldsymbol{y}}$，*goal*，*spread* 分别为输入向量矩阵、目标向量矩阵、均方误差和 RBF 的分布。

3. 用 sim 函数对时间序列进行预测

格式：b=sim(net，$\boldsymbol{a}$)

其中，$\boldsymbol{a}$，b 分别为待评价时间序列的输入向量和用 RBF 网络对时间序列进行计算的预测值。

以上 3 步构成时间序列预测的 Matlab6.5 环境下的 RBF 网络模型。

9.4.3 环境应用

例 9.4 海洋冰情时间序列是海洋灾害管理的基本资料之一，对其进行有效预测，可为减轻海洋冰灾损失提供重要的理论指导。因冰情序列受众多不确定性因素影响，在时序上常常表现出弱相依性、突变性和随机性等复杂的非线性特征，至今对此进行有效描述的数学模型仍不很成熟。为了说明上述模型的有效性，现利用表 9.7 中 1966～1993 年度冰情等级资料序列$\{x^*(i), i=1\sim27\}$（杨晓华等，1999；余加艾等，1995）来建立 RBF 网络冰情预测模型。表 9.7 中，1966 表示 1966～1967 年度，1993 表示 1993～1994 年度，余类推。

计算该序列前 6 阶自相关系数值 $R(k)$ 和与之相应的式(9.1)右边上、下限 $R_2(k)$、$R_1(k)$ 值，结果见表 9.8，其中置信水平取 70%。表 9.8 显示，只有 $R(1)$、$R(3)$、$R(4)$ 的相依性在置信水平 70% 的条件下是显著的，故这里以最大相依性延迟步数 $m=4$，作为 RBF 网络模型输入向量的维数。

表 9.7 某海洋冰情等级序列实测值和各模型的拟合结果与预测结果　单位：冰级

年度	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
实测值	3.00	4.50	5.00	3.00	3.50	3.00	1.00	3.00	1.50	1.50	4.50	2.50	2.50	3.00
TAR计算值					3.05	2.60	1.77	2.65	2.45	1.65	2.93	2.63	2.87	3.62
绝对误差*					0.45	0.40	−0.77	0.35	−0.95	−0.15	1.57	−0.13	−0.37	−0.62
RBF 方法					3.50	3.00	1.00	3.00	1.50	1.50	4.50	2.50	2.50	3.00
绝对误差**					0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
年度	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
实测值	2.50	2.50	2.00	3.00	3.50	3.00	3.00	2.00	1.50	3.00	1.50	1.50	1.50	1.50
TAR计算值	1.87	2.66	2.79	2.33	2.78	2.76	3.13	2.80	2.22	2.53	2.42	2.07	2.83	1.87
绝对误差*	0.63	−0.16	−0.79	0.67	0.72	0.24	−0.13	−0.80	−0.72	0.47	−0.92	−0.57	−1.33	−0.37
RBF 方法	2.50	2.50	2.00	3.00	3.50	3.00	3.00	2.00	1.50	3.00	1.50	1.50	1.50	1.76
绝对误差**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	−0.26

注：* 表示 TAR 的绝对误差，** 表示 RBF 方法的绝对误差。

表 9.8 某海洋冰情等级序列自相关系数及其上、下限值(置信水平 70%)

k	1	2	3	4	5	6
$R_1(k)$	−0.238	−0.244	−0.249	−0.256	−0.262	−0.269
$R(k)$	0.251	0.105	0.217	−0.278	−0.100	−0.110
$R_2(k)$	0.162	0.164	0.166	0.169	0.171	0.174

取表 9.7 中 1970～1993 年度冰情等级作为训练样本，用 Matlab6.5 环境下的 RBF 网络法对 1993～1994 年度冰情等级的预测结果参见表 9.7。RBF 网络计算过程如下：

1. 用式(9.2)建立 $n-m$ 组训练样本的输入、输出向量。

$$\bar{x}=[x^{m+1},\ x^{m+2},\ \cdots,\ x^{n}],\quad \bar{y}=[y^{m+1},\ y^{m+2},\ \cdots,\ y^{n}]$$

这里，$n=27$，$m=4$。

2. 设计 RBF 网络。令 $g=0.000\ 01$，$s=1$

$$\text{net}=\text{newrb}(\bar{x},\ \bar{y},\ g,\ s)$$

其中，newrb 为径向基网络设计函数，g 为训练精度，s 为径向基层的散布常数。

3. 由 $t=\text{sim}(\text{net},\ \bar{x})$，可得网络的训练结果。

sim 为模拟函数，由 plot(h，$\bar{y}$，'k×'，h，t，'k○')可得网络训练的图形输出结果，如图 9-19。这里 h 代表各冰情等级所对应的年度序号，$h=(5,\ 6,\ \cdots,\ 27)$。$\bar{y}$ 代表各年度冰情等级的目标值，用“×”表示。t 代表各年度冰情等级的计算值，用“○”表示。从图 9-19 可以看出，计算误差为 0。

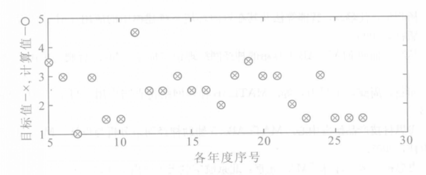


图 9-19 网络的训练结果

由 $\boldsymbol{a}=(3.0,\ 1.5,\ 1.5,\ 1.5)$ 和 $b=\text{sim}(\text{net},\ \boldsymbol{a})$，可得到 1993～1994 年度冰情等级的预测结果 $b=1.756\ 6$，详见表 9.7。

在表 9.7 中 1993～1994 年度为试报，其余为历史预报。从表 9.7 可知，在 23 次历史预报中，每次绝对误差均为 0.00，训练样本合格率为 100%，而试报绝对误差小于 0.30，较 TAR 模型的精度有了显著提高。RBF 网络模型虽然仅利用海洋冰情等级时序延迟 1 步、延迟 2 步、延迟 3 步和延迟 4 步的相依信息，但由于 RBF 网络可以描述该时序非线性动态系统，其拟合精度和预测精度都是令人满意的。

【思考题 9】

1. 试述 BP 神经网络原理。
2. BP 神经网络采用的激发函数为什么必须是连续可导的？
3. 给出 BP 算法的基本步骤及计算框图。
4. 试述 BP 算法的优缺点。
5. 试述 RBF 神经网络原理及计算框图。
6. 试述 RBF 神经网络的优缺点。
7. 试比较 BP 与 RBF 神经网络的性能。
8. 试用 BP 神经网络解决一个实际环境问题。
9. 试用 RBF 神经网络解决一个实际环境问题。

【参考文献】

[1] 杨晓华，沈珍瑶. 智能算法及其在资源环境系统建模中的应用 [M]. 北京：北京师范大学出版社，2005.

[2] 丛爽. 面向 MATLAB 工具箱的神经网络理论与应用 [M]. 合肥：中国科学技术大学出版社，2003.

[3] 闻新，周露，王丹力，等. MATLAB 神经网络仿真与应用 [M]. 北京：科学出版社，2003.

[4] 飞思科技产品研发中心. MATLAB6.5 辅助神经网络分析与设计 [M]. 北京：电子工业出版社，2003.

[5] 史忠植. 神经计算 [M]. 北京：北京航空航天大学出版社，1995.

[6] 金菊良，丁晶. 遗传算法及其在水科学中的应用 [M]. 成都：四川大学出版社，2000.

[7] 陈守煜. 中长期水文预报综合分析理论模式与方法 [J]. 水利学报，1997 (8)：15-21.

[8] 郭劲松，霍国友，龙腾锐. BOD-DO耦合人工神经网络水质模拟的研究 [J]. 环境科学学报，2001，21(2)：140-143.

[9] 余加艾，刘钦政. 利用灰色系统方法预测冰情 [J]. 海洋环境科学，1995，14(4)：70-75.

[10] 杨晓华，金保明，金菊良，等. 门限自回归模型在海洋冰情预测中的应用 [J]. 灾害学，1999，14(4)：1-6.

[11] 金菊良，丁晶，魏一鸣. 基于遗传算法的门限自回归模型在海温预测中的应用 [J]. 海洋环境科学，1999，18(3)：1-6.

[12] 马玉梅，高静宇，王清华. 基于人工神经网络的赤潮预测模型 [J]. 海洋预报，2007，24(1)：38-44.

第 10 章　环境空间统计分析

环境信息一般指来自环境保护和社会相关部门，采用一定的技术手段或方法采集的反映环境空间系统里环境质量状况、污染物排放、自然生态和环境保护工作等各种数据资料的总体集合。可以被认为是一种已被加工为特定形式的环境数据，或是一组表示数量、行动或目标的可鉴别的符号，它可以是数字、字母或符号，也可以是图形、图像或声音等，并可按使用目的组织成结构型数据库或非结构型数据库。环境信息有一个非常突出和重要的特性即空间性。据统计，环境信息 85%以上都与空间位置有关，可以把具有空间属性的环境信息称为环境空间信息，它是具体描述地球环境中实体的空间特征、属性特征和时间特征的数据集合。常见的环境空间信息有污染源分布、监测站点分布、环境质量的空间分异特征等。环境空间信息来源繁多，结构多样，应用领域非常广泛。

本章的主要内容是：

- 信息与数据；
- 环境空间信息；
- 环境空间统计分析；
- 环境空间主成分分析。

10.1　环境空间信息概述

信息是近代科学的一个专门术语，已广泛地应用于社会各个领域。关于信息有各种不同的定义，狭义信息是指人们获得信息前后对事物认识的差别；广义信息是指主体(人、生物和机器)与外部客体(环境、其他人、生物和机器)之间相互联系的一种形式，是主体和客体之间一切有用的消息和知识，是表征事物特征的一种普遍形式。总之，信息是向人们或机器提供关于现实世界各种事实的知识，是数据、消息中所包含的意义，它不随载体的物理形式的各种改变而改变。

数据是通过数字化或直接记录下来的可以被鉴别的符号，不仅数字是数据，而且文字、符号和图像也是数据。数据是用以载荷信息的物理符号，在计算机化的环境信息系统中，数据的格式往往和具体的计算机系统有关，随载荷它的物理设备的形式而改变。数据只有对实体行为产生影响时才成为信息。例如同样的数据“1”和“0”，当用来表示某一种实体在某个地域内存在与否时，它就提供了

有(用“1”表示)无(用“0”表示)的信息；在绘图矩阵中表示绘线或不绘线时，它就提供落笔抬笔的信息。

信息与数据虽然有词义上的差别，但两者是不可分离的，即信息是数据的内涵，数据是信息的表达。环境信息的建立和进行，就是信息(或数据)按一定方式流动的过程。通常情况下，并不严格地区分“信息”和“数据”两个术语。

环境信息是指表征环境系统诸要素的数量、质量、分布特征、相互关系和变化规律的数字、文字、图像和图形等的总称，用文字、数字、符号、图像等不同形式定性、定量、定位、定时，可视化地全面表征环境的这些属性特征。环境信息表征了有关环境实体的性质、特征和变化状态，是对表达环境特征与现象之间关系的环境数据的解释。

环境空间信息是指具有空间坐标的环境信息，强调的是具有空间位置的信息，更强调信息中所蕴含的内容和属性。这些信息(数据)可以是图形、图像、文字、表格和数字等，通过数字化仪、扫描仪、键盘、磁带机或其他系统通讯输入计算机，是计算机所表达的现实世界经过模型抽象的实质性内容。

10.1.1 环境空间信息特征

环境空间信息除了具有一般信息的特征之外，还具有一些区别于其他信息的特性。构成环境空间信息的特征主要有：

1. 空间性

空间性是环境空间信息最主要的特性。环境空间信息描述了环境空间物体的位置、形态，甚至需要描述物体的空间拓扑关系。例如描述一条河流污染，一般数据侧重于河流的污染物含量等，而环境空间信息则要附加污染源的位置等和空间位置有关的信息。复杂一点的还要处理河流与流域内城市间的距离、方位等空间关系。空间性是空间数据区别于其他数据的标志特征。

2. 抽象性

环境空间信息描述的是现实世界中地物的污染特征，自然界中地物非常复杂，必须经过抽象处理。不同主题的空间数据库，人们所关心的内容也有差别，所以环境空间信息的抽象性还包括人为地取舍数据。抽象性还使数据产生多语义问题。在不同的抽象中，同一自然地物的表示可能会有不同的语义，如河流既可以被抽象为水系要素，也可以被抽象为行政边界，如省界，县界等。

3. 多尺度与多态性

不同的观察尺度具有不同的比例尺和不同的精度，同一地物在不同的情况下就会有形态差异。

4. 多时空性

环境空间数据具有很强的时空特性。一个环境空间信息系统中的数据源既有同一时间不同空间的数据系列，也有同一空间不同时间序列的数据。不仅如此，环境空间信息系统会根据系统需要而采用不同尺度对环境空间进行表达。环境空间数据是包括不同时空和不同尺度数据源的集成。

10.1.2 环境空间信息种类

若按照环境空间信息的内容和特性分类，常见的环境空间信息有以下五种：

(1)环境监测数据。主要包括空气质量和废气监测数据、降水监测数据、地表水和废水监测数据、土壤底质固体废物监测数据、生物监测数据、噪声振动监测数据、森林生态系统、荒漠生态系统、农业生态系统监测数据、生态破坏监测数据、化学污染监测数据、淡水监测数据、湿地生态监测数据、海洋生态监测数据等。

(2)工业污染与防治数据。包括工业污染企业基本情况、工业污染物排放情况、固体废弃物排放情况、工业污染治理设施情况、工业企业在建污染治理项目情况等。

(3)生活及其他污染与防治数据。包括生活污水排放情况、城市污水处理情况、生活废气排放情况、城市垃圾处理情况、规模化畜禽养殖场污染排放及治理情况等。

(4)自然生态环境保护数据。包括自然保护区建设情况、野生动植物保护情况、生态示范区建设情况、农村环境污染及治理情况等。

(5)环境管理数据。包括法律法规、环保年度计划执行、跨世纪绿色工程规划执行情况、建设项目环境影响评价、三同时(同时设计、同时施工、同时投入使用)执行情况、环境科技工作情况、环保产业情况、环保系统自身建设情况等。

10.1.3 环境空间信息来源

1. 理论来源

信息是用文字、数字、符号、语言、图像等介质来表示事件、事物、现象等的内容、数量或特征，从而向系统(人们)提供关于现实世界新的事实和知识。信息具有客观性、实用性、可传输性和共享性等特征，它是事物特征及事物之间相互联系的抽象反映。这种反映能被人们认识和理解并作为知识来识别事物，从而达到认识世界、改造世界的目的。因此，信息可以作为生产、建设、经营、管

理、分析和决策的依据，成为了当今社会和未来社会最重要的战略资源。

而环境信息属于空间信息，其位置的识别是与数据联系在一起的。环境信息的这种定位特征是通过经纬网建立的环境坐标来实现空间位置识别的。环境信息还具有多维结构的特征，即在二维空间的基础上实现多专题的第三维结构，而各个专题型、实体型信息之间的联系是通过属性码进行的。这就为环境系统综合研究提供了可能，也为环境系统多层次的分析和信息的传输与筛选提供了方便。环境信息的时序特征十分明显，因此可以按照时间尺度对环境信息进行划分。环境信息的这种动态变化的特征，一方面要求环境信息的获取要及时，并定期更新；另一方面要从其变化过程中研究其变化规律，从而做出对环境需求的预测，为科学决策提供依据。认识环境信息的这种区域性、多层次性和动态变化的特征对建立环境信息系统，实现人口、资源、环境等的综合分析、管理、规划和决策具有重要意义。

2. 技术来源

"信息"这一词汇作为科技术语，20世纪60年代初期开始在科技文献上出现，直到20世纪80年代初期，随着微型计算机的普及，才普遍被人们接受。从20世纪80年代开始，各个学科、各个领域主动改变研究方式，引领计算机向着信息化方向发展。有资料表明，现今的计算机只有24%的工作量用于单纯的计算，而76%的工作量用作信息的存取、检索与处理。

同一类信息数据在计算机系统安排下形成数据库。大量的甚至海量的数据组织在一起有一个管理的问题，这就需要有相应的数据库管理技术。计算机的发明催生了数据库，计算机软、硬件的更新，推动着数据库技术变革。最初的数据库的概念只是指一个数据文件或一个数据表格，计算机系统对文件或表格的容量、数据格式的限制很多，修改、编辑也较困难。随着系统对数据库允许的数据容量剧增，管理能力大幅度增强，直至最后产生关系数据库，即数据库管理系统在指令的驱动下，可以对数据文件集合进行一致性的数据变更维护，即对有相互关联关系的多个数据文件表格进行自动一次性联动变更。

计算机数据库为人们从大量的信息数据库中快速检索、提取信息提供了极为便利的条件。但事物是相互联系、相互作用的，仅静止地检索出某一条孤立的信息，往往会使这一信息的价值受到一定的限制，比如，我们如果仅从环境信息数据库中检索出一个地区的环境类型、地区的降雨量、地区的植被分布，这些信息对于环境的科学管理仍然是不够的。我们需要的是将这些信息加以综合考虑。从事物的相互联系又相互作用的观点出发，将相互作用的模式分析推导出来，以计算机能够接受的形式交付计算机按照这种模式分析、处理信息，这就是信息系统所要完成的任务。

经过近几十年的发展，信息系统已经成为一个独立的学科门类，并有多个分支，环境信息系统就是其中的典型。

3. 社会来源

信息是决策和管理的基础，环境信息是环保等部门进行决策和管理的主要依据。环保部门的决策能力和办事效率在很大程度上取决于信息工作的水平和质量。而信息准确可靠、信息之间具有可比性，在很大程度上依赖于信息本身的标准化、规范化程度。

环境保护是我国的一项基本国策。随着我国环境保护事业的发展，环境管理工作不断深化，信息化已成为提高环境管理与决策水平的重要技术基础。环境信息的采集面越来越广，人们对环境信息的需求量也越来越大，创建科学、合理的环境信息系统已成为必需。

10.2 环境空间统计分析

空间统计学(spatial statistics)又称地统计学(geostatistics)，是近几十年来发展起来的一种新的分析方法，它包括空间结构分析、克立格分析、空间自相关分析以及空间模拟等技术，用于分析具有空间坐标的变量的空间特征，并可进行过程模拟以及空间插值等。

空间统计学是以区域化变量理论(theory of regionalized variable)为基础，以变差函数(variogram)为基本工具来研究那些分布于空间并呈现出一定的随机性和结构性的自然现象的科学(A. G. 儒尔奈耳等，1982；王仁铎等，1988；孙洪泉，1990)。显然，凡是要研究某些变量(或特征)的空间分布特性并对其进行最优估计，或要模拟所研究对象的离散性、波动性或其他性质时都可应用空间统计学的理论与方法。

空间统计学是数学地质领域中一门发展迅速且有着广泛应用前景的新兴科学。空间统计学的基本思想从 20 世纪 50 年代初开始提出，经过广大数学地质工作者、空间统计学工作者、矿山地质和采矿设计专家及其他空间统计学应用者和爱好者的不断努力，现在已经形成了一套独立的理论体系，成为数学地质中比较活跃的一个分支(M. Guarascio 等，1975；F. P. 阿格特伯格，1980；M. 戴维，1989)。空间统计学在国内外诸多领域的生产实践中表明，除了在找矿勘探、矿体圈定、储量计算、采矿设计、矿山生产及地学科研等方面具有明显的优越性外，在石油地质、生物学、生态学、岩石学、地球化学、地震地质、海洋地质、农业、水文、古气候、古地理、气象学、遥感地质、环境、林业、医学等许多方面都有成功应用的实例(於崇文等，1980；侯景儒等，1982；侯景儒等，1993；

王政权，1999)。因此，在不到 50 年的研究和实践中，空间分析的应用已被扩展到分析各种自然现象的空间异质性(spatial heterogeneity)和空间格局(spatial pattern)。环境空间统计分析就是应用空间统计学的理论和方法处理环境空间信息的过程。

10.2.1　区域化变量

当一个变量呈现为空间分布时，就称之为区域化变量(regionalized variable)。这种变量常常反映某种空间现象的特征，用区域化变量来描述的现象称之为区域化现象(I. Clark，1981)。例如，地质学、地理学、水文学、土壤学、生态学中的许多变量都具有空间分布的特点，这些变量实质上都是区域化变量。

区域化变量，亦称区域化随机变量，G. Matheron 将它定义为以空间点 $\boldsymbol{x}$ 的三个直角坐标 x_u，x_v，x_w 为自变量的随机场 $Z(\boldsymbol{x})=Z(x_u,\ x_v,\ x_w)$。区域化随机变量与普通随机变量不同，普通随机变量的取值符合某种概率分布，而区域化随机变量则根据其在一个场内的位置不同而取值。也就是说，区域化随机变量是普通随机变量在一个场内确定位置上的特定取值，它是与位置有关的随机函数。在对所研究的空间对象进行一次抽样或随机观察后就得到它的一个 $Z(\boldsymbol{x})$，它是一个普通的三元实值函数，或者说是空间的点函数。因此，区域化变量具有两方面的含义，即观测前 $Z(\boldsymbol{x})$是一个随机变量，观测后则是一个普通的空间三元函数值或空间点函数值。

区域化变量 $Z(\boldsymbol{x})$具有两个最显著、最重要的特征，即随机性和结构性。正是这两种性质使区域化变量在研究自然现象的空间结构和空间过程方面具有独特的优势。首先，区域化变量是一个随机函数，它具有局部的、随机的、异常的性质；其次，区域化变量具有一般的或平均的结构性质，即变量在点 $\boldsymbol{x}$ 与偏离空间距离为 $\boldsymbol{h}$ 的点 $\boldsymbol{x}+\boldsymbol{h}$ 处的数值 $Z(\boldsymbol{x})$与 $Z(\boldsymbol{x}+\boldsymbol{h})$具有某种程度的自相关，这种自相关依赖于两点间的距离 h 及变量特征，这就体现了其结构性。此外，区域化变量还具有空间的局限性、不同程度的连续性和不同程度的各向异性等特征。

由于区域化变量具有上述特点，需要有一种合适的函数或模型来描述，这种函数和模型既能兼顾到区域化变量的随机性，又能反映它的结构性。这可以通过描述空间变异性的空间协方差函数和变差函数来实现。

10.2.2 协方差函数

1. 协方差函数的概念

区域化随机变量之间的差异，可以用空间协方差来表示。协方差又叫做半方差，是空间统计学中的关键概念。

在概率论中，随机向量（x，y）的协方差被定义为：

$$\mathrm{Cov}(x, y)=E[(x-E(x))(y-E(y))] \tag{10.1}$$

区域化变量 $Z(\boldsymbol{x})=Z(x_u, x_v, x_w)$在空间点 $\boldsymbol{x}$ 和 $\boldsymbol{x}+\boldsymbol{h}$ 处的两个随机变量 $Z(\boldsymbol{x})$和 $Z(\boldsymbol{x}+\boldsymbol{h})$的二阶混合中心矩定义为 $Z(\boldsymbol{x})$的自协方差函数，即：

$$\mathrm{Cov}[Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})]=E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})]-E[Z(\boldsymbol{x})]E[Z(\boldsymbol{x}+\boldsymbol{h})] \tag{10.2}$$

区域化变量 $Z(\boldsymbol{x})$的自协方差函数，也简称为协方差函数。一般来讲，它是一个依赖于空间点 $\boldsymbol{x}$ 和向量 $\boldsymbol{h}$ 的函数。

2. 协方差函数的计算公式

设 $Z(\boldsymbol{x})$为区域化随机变量，并满足二阶平稳假设，即随机函数 $Z(\boldsymbol{x})$的空间分布规律不因位移而改变，$\boldsymbol{h}$ 为两样本点空间分隔距离或距离滞后，$Z(\boldsymbol{x}_i)$为 $Z(\boldsymbol{x})$在空间位置 $\boldsymbol{x}_i$ 处距离偏移 $\boldsymbol{h}$ 的实测值($i=1, 2, \cdots, N(\boldsymbol{h})$)，根据协方差函数的定义，可得协方差函数的计算公式为：

$$C(\boldsymbol{h})=\frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}[Z(\boldsymbol{x}_i)-\overline{Z}(\boldsymbol{x}_i)][Z(\boldsymbol{x}_i+\boldsymbol{h})-\overline{Z}(\boldsymbol{x}_i+\boldsymbol{h})] \tag{10.3}$$

式中，$N(\boldsymbol{h})$是分隔距离为 $\boldsymbol{h}$ 时的样本点对总数，$\overline{Z}(\boldsymbol{x}_i)$和 $\overline{Z}(\boldsymbol{x}_i+\boldsymbol{h})$分别为 $Z(\boldsymbol{x}_i)$和 $Z(\boldsymbol{x}_i+\boldsymbol{h})$的样本平均数，即：

$$\overline{Z}(\boldsymbol{x}_i)=\frac{1}{N}\sum_{i=1}^{N}Z(\boldsymbol{x}_i) \tag{10.4}$$

$$\overline{Z}(\boldsymbol{x}_i+\boldsymbol{h})=\frac{1}{N}\sum_{i=1}^{N}Z(\boldsymbol{x}_i+\boldsymbol{h}) \tag{10.5}$$

在式(10.4)～(10.5)中，N 为单元样本数。一般情况下，$\overline{Z}(\boldsymbol{x}_i)\neq\overline{Z}(\boldsymbol{x}_i+\boldsymbol{h})$(特殊情况下可以认为近似相等)。若 $\overline{Z}(\boldsymbol{x}_i)=\overline{Z}(\boldsymbol{x}_i+\boldsymbol{h})=m$(常数)，则式(10.3)可以改写为：

$$C(\boldsymbol{h})=\frac{1}{N(\boldsymbol{h})}\sum_{i=1}^{N(\boldsymbol{h})}[Z(\boldsymbol{x}_i)Z(\boldsymbol{x}_i+\boldsymbol{h})]-m^2 \tag{10.6}$$

式中，m 为样本平均数，可由一般算术平均数公式求得，即：

$$m=\frac{1}{N}\sum_{i=1}^{n}Z(\boldsymbol{x}_i) \tag{10.7}$$

10.2.3 变差函数

10.2.3.1 二阶平稳假设和本征假设

1. 平稳假设(stationarity assumption)

(1) 严格的平稳假设：假设区域化变量 $Z(\boldsymbol{x})$ 的任意 n 维分布函数均不因空间点 $\boldsymbol{x}$ 发生位移 $\boldsymbol{h}$ 而改变，即：

$$
\begin{aligned}
&F_{x_1,x_2,\cdots,x_n}(z_1, z_2, \cdots, z_n)\\
&=P\{Z(\boldsymbol{x}_1)<z_1, Z(\boldsymbol{x}_2)<Z_2, \cdots, Z(\boldsymbol{x}_n)<z_n\}\\
&=P\{Z(\boldsymbol{x}_1+\boldsymbol{h})<z_1, Z(\boldsymbol{x}_2+\boldsymbol{h})<z_2, \cdots, Z(\boldsymbol{x}_n+\boldsymbol{h})<z_n\} \quad (10.8)\\
&=F_{x_1+h,x_2+h,\cdots,x_n+h}(z_1, z_2, \cdots, z_n), \ \forall n, \forall \boldsymbol{h}, \forall \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n
\end{aligned}
$$

则称该区域化变量 $Z(\boldsymbol{x})$ 为平稳性随机函数。确切地说，无论位移向量 $\boldsymbol{h}$ 多大，两个 n 维区域化变量 $\{Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2), \cdots, Z(\boldsymbol{x}_n)\}$ 和 $\{Z(\boldsymbol{x}_1+\boldsymbol{h}), Z(\boldsymbol{x}_2+\boldsymbol{h}), \cdots, Z(\boldsymbol{x}_n+\boldsymbol{h})\}$ 具有相同的分布律。然而这种假设条件性太强，至少要求 $Z(\boldsymbol{x})$ 的各阶矩均存在，实际上很难满足，且也不好验证，故实用上不采用这种假设。在线性空间统计学研究中，主要研究方差，为了统计推断的需要，我们只需假设 $Z(\boldsymbol{x})$ 的一、二阶矩存在且平稳就够了。故在实际应用中常用另一种弱平稳假设，或称为二阶平稳假设(second order stationarity assumption)。

(2) 二阶平稳假设：当区域化变量 $Z(\boldsymbol{x})$ 满足下面两个条件时，则称该区域化变量为二阶平稳的。

① 在整个研究区内，区域化变量 $Z(\boldsymbol{x})$ 的数学期望对任意 $\boldsymbol{x}$ 存在且等于常数，即：

$$E[Z(\boldsymbol{x})]=m(\text{常数}), \ \forall \boldsymbol{x} \tag{10.9}$$

② 在整个研究区内，区域化变量 $Z(\boldsymbol{x})$ 的协方差函数存在且平稳(即只依赖于基本步长 $\boldsymbol{h}$，而与 $\boldsymbol{x}$ 无关)，用式子表达，即：

$$
\begin{aligned}
\mathrm{Cov}[Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})]&=E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})]-E[Z(\boldsymbol{x})]E[Z(\boldsymbol{x}+\boldsymbol{h})]\\
&=E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})]-m^2 \overset{\text{记为}}{=} C(\boldsymbol{h})\\
&=E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})]-m^2=C(\boldsymbol{h}), \ \forall \boldsymbol{x}, \ \forall \boldsymbol{h}
\end{aligned} \tag{10.10}
$$

当 $\boldsymbol{h}=\boldsymbol{0}$ 时，上式变为：

$$D[Z(\boldsymbol{x})]=C(\boldsymbol{0}), \ \forall \boldsymbol{x} \tag{10.11}$$

此式说明：方差函数也存在，且为常数 $C(\boldsymbol{0})$。

这说明协方差平稳意味着方差和变差函数平稳，从而有关系式：

$$C(\boldsymbol{h})=C(\boldsymbol{0})-\gamma(\boldsymbol{h}) \tag{10.12}$$

同时还说明，在二阶平稳假设条件下，协方差函数和变差函数都表示相距为 $\boldsymbol{h}$ 的两个变量 $Z(\boldsymbol{x})$ 和 $Z(\boldsymbol{x}+\boldsymbol{h})$ 之间的自相关特性，这时它们两个是等效的函数。这样就可以定义出第三个空间函数，即空间相关函数(correlogram)，简称相关函数，记为 $\rho(\boldsymbol{h})$，即：

$$\rho(\boldsymbol{h})=\frac{C(\boldsymbol{h})}{C(\boldsymbol{0})}=1-\frac{\gamma(\boldsymbol{h})}{C(\boldsymbol{0})} \tag{10.13}$$

在实际工作中，有时连二阶平稳假设的要求也不能满足(如协方差函数或方差函数不存在等)。例如，一些自然现象和随机函数，它们具有无限离散性，即无协方差及先验方差，但却有变差函数，这时，我们可以放宽条件，如只考虑品位的增量而不考虑品位本身，于是导致本征假设，即内蕴假设。

2. 本征假设(intrinsic assumption，内蕴假设)

当区域化变量 $Z(\boldsymbol{x})$ 的增量 $[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]$ 满足下列两个条件时，称其为满足本征假设，或简单地说它是本征的：

(1) 在整个研究区内，区域化变量 $Z(\boldsymbol{x})$ 的增量 $[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]$ 的数学期望对任意的 $\boldsymbol{x}$ 和 $\boldsymbol{h}$ 都存在且等于零，即：

$$E[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]=0,\ \forall \boldsymbol{x},\ \forall \boldsymbol{h} \tag{10.14}$$

(2) 在整个研究区内，区域化变量 $Z(\boldsymbol{x})$ 的增量 $[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]$ 的方差函数存在且平稳，即：

$$\begin{aligned}D[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]&=E[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]^2\\&=2\gamma(\boldsymbol{x},\ \boldsymbol{h})\\&=2\gamma(\boldsymbol{h}),\ \forall \boldsymbol{x},\ \forall \boldsymbol{h}\end{aligned} \tag{10.15}$$

即要求 $Z(\boldsymbol{x})$ 的变差函数 $\gamma(\boldsymbol{h})$ 存在且平稳。

本征假设可以理解为：区域化变量 $Z(\boldsymbol{x})$ 的增量 $[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]$ 只依赖于分割它们的向量 $\boldsymbol{h}$(模和方向)而不依赖于 $\boldsymbol{x}$ 的具体位置，这样，被向量 $\boldsymbol{h}$ 分割的每一对数据 $[Z(\boldsymbol{x}),\ Z(\boldsymbol{x}+\boldsymbol{h})]$ 可以看成是一对随机变量 $[Z(\boldsymbol{x}_1),\ Z(\boldsymbol{x}_2)]$ 的一个不同实现。

3. 二阶平稳假设与本征假设之比较

二阶平稳假设与本征假设比较的总结论是：二阶平稳假设较强，本征假设较弱。满足二阶平稳假设的区域化变量必定满足本征假设；满足本征假设的区域化变量，却不见得满足二阶平稳假设。故满足本征假设的区域化变量要广一些，多一些。

4．准二阶平稳假设和准本征假设

在实际应用中，往往遇到这样的情况，即区域化变量 $Z(\boldsymbol{x})$ 在整个区域内并不满足二阶平稳假设(或本征假设)，但在有限大小的邻域(例如，以 $\boldsymbol{x}$ 点为中心，以 a 为半径的球或圆)内是二阶平稳(或本征)的，则称此区域化变量 $Z(\boldsymbol{x})$ 是准二阶平稳(或准本征)的。

这种假设虽是一种折衷方案，但在现实中能满足的往往就是这种假设，而且在实际空间统计学计算中这种假设也够用了。不过这种假设涉及到有限邻域的大小应如何确定的问题。邻域确定大了，往往不易满足准二阶平稳(或准本征)假设条件；邻域确定小了，虽能满足假设条件，但邻域内信息数据点就少了，又不利于进行统计判断。故在确定合适的邻域大小时要兼顾上述两个方面。

以后我们在讨论线性平稳空间统计学时，都至少假定 $Z(\boldsymbol{x})$ 满足准二阶平稳假设条件或准本征假设条件。

有了这种假设，我们便可根据 n 对 $Z(\boldsymbol{x}_i)$ 和 $Z(\boldsymbol{x}_i+\boldsymbol{h})$ $(i=1,\ 2,\ \cdots,\ n)$ 的数值，通过求某种平均数的办法来估计变差函数值了。

10.2.3.2　变差函数

1．定义

将环境空间信息看作成随空间位置 $\boldsymbol{x}$ 而变化的区域化变量 $Z(\boldsymbol{x})$(为讨论问题方便不妨设 $Z(\boldsymbol{x})$ 定义在一维坐标轴上)，那么，当空间点 $\boldsymbol{x}$ 在一维 $\boldsymbol{x}$ 轴上变化时，区域化变量 $Z(\boldsymbol{x})$ 在点 $\boldsymbol{x}$ 和 $\boldsymbol{x}+\boldsymbol{h}$ 处的值 $Z(\boldsymbol{x})$ 与 $Z(\boldsymbol{x}+\boldsymbol{h})$ 之差的方差的一半定义为区域化变量 $Z(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 轴方向上的变差函数，记作 $\gamma(\boldsymbol{x},\ \boldsymbol{h})$。即：

$$\gamma(\boldsymbol{x},\ \boldsymbol{h})=\frac{1}{2}D[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})] \tag{10.16}$$

根据协方差函数的理论，变差函数可以展开为：

$$\begin{aligned}\gamma(\boldsymbol{x},\ \boldsymbol{h})&=\frac{1}{2}D[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]\\&=\frac{1}{2}E[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]^2-\frac{1}{2}\{E[Z(\boldsymbol{x})]-E[Z(\boldsymbol{x}+\boldsymbol{h})]\}^2\end{aligned} \tag{10.17}$$

在实际的空间统计学研究中，多要作一些假设。通常是作二阶平稳假设或作本征假设。在这两种假设下均有：

$$E[Z(\boldsymbol{x}+\boldsymbol{h})]=E[Z(\boldsymbol{x})],\ \forall \boldsymbol{h} \tag{10.18}$$

因此，式(10.17)就可以简化为：

$$\gamma(\boldsymbol{x},\ \boldsymbol{h})=\frac{1}{2}E[Z(\boldsymbol{x})-Z(\boldsymbol{x}+\boldsymbol{h})]^2 \tag{10.19}$$

这是空间统计学中最常用的基本公式之一。

从式(10.19)中可以看出，$\gamma(\boldsymbol{x}, \boldsymbol{h})$一般是依赖于$\boldsymbol{x}$和$\boldsymbol{h}$两个自变量的。当变差函数$\gamma(\boldsymbol{x}, \boldsymbol{h})$仅依赖于$\boldsymbol{h}$(基本步长或基本滞后)而与位置$\boldsymbol{x}$无关时，则可把变差函数$\gamma(\boldsymbol{x}, \boldsymbol{h})$写成$\gamma(\boldsymbol{h})$，即：

$$\gamma(\boldsymbol{h})=\frac{1}{2}E[Z(\boldsymbol{x})-E(\boldsymbol{x}+\boldsymbol{h})]^2 \tag{10.20}$$

此时，以$\boldsymbol{h}$为横坐标，以$\gamma(\boldsymbol{h})$值为纵坐标作出的图形就叫变差图。故变差函数与变差图严格说来，还是有区别的。但当变差函数$\gamma(\boldsymbol{x}, \boldsymbol{h})$不依赖于$\boldsymbol{x}$时，这两者就是一样的，只不过一个代表函数关系式，另一个表示其函数的图形罢了。

如果$Z(\boldsymbol{x})$是定义在二维(或三维)空间中的区域化变量，则$\boldsymbol{x}$是二维(或三维)空间中的点，$\boldsymbol{h}$是二维(或三维)空间中的向量(此时，x，h本应写成$\boldsymbol{x}$，$\boldsymbol{h}$，为了简化，在不致发生混淆处，就写成标量形式)。此时，就要考虑二维(或三维)变差函数了。

2. 实验变差函数

在实际工作中，要对区域化变量$Z(x)$做变异性分析，通常是先求出实验变差函数，然后再用理论模型拟合，得到最终的变差函数公式。对于离散点的情况，由于有了(准)二阶平稳假设或(准)本征假设，我们可以把在x轴上相隔为h的$N(h)$对点x_i和$x_i+h(i=1, 2, \cdots, N(h))$处的$N(h)$对观测值$Z(x_i)$和$Z(x_i+h)(i=1, 2, \cdots, N(h))$看成是$Z(x)$和$Z(x+h)$的$N(h)$对实现。其实验变差函数的基本公式为：

$$\gamma^*(h)=\frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(x_i)-Z(x_i+h)]^2 \tag{10.21}$$

这样，对于不同的空间分隔距离h，根据式(10.21)可计算出相应的$\gamma^*(h)$值来。这就是计算实验变差函数的最基本的公式。经计算后，得出诸对h，$\gamma^*(h)$值，在$h-\gamma^*(h)$直角坐标上标出诸点$(h, \gamma^*(h))$来，再将相邻各点用直线段连接起来，就得到实验变差函数图(或称实验变差图)。这样的曲线图可以直接地展示参数区域化变量$Z(x)$的空间变异特点，是空间变异分析和结构分析的有效工具。

3. 变差函数的理论模型

由区域化变量理论和变差函数的性质可知，实际上，理论变差函数模型是未知的，往往要从有效的空间取样数据中去估计，对各种不同的h值可以计算出一系列$\gamma(h)$值。到目前为止，空间统计学将这些模型分为三大类：第一类是有基台值模型，包括球状模型、指数模型、高斯模型、线性有基台值模型和纯块金效应模型；第二类是无基台值模型，包括幂函数模型、线性无基台值模型、抛物线

模型；第三类是孔穴效应模型。下面有代表性地介绍几种常见的变差函数理论模型。

（1）纯块金效应模型。其一般公式为：

$$\gamma(h)=\begin{cases}0 & (h=0)\\ c_0 & (h>0)\end{cases} \tag{10.22}$$

式中，$c_0>0$，为先验方差。该模型相当于区域化变量为随机分布，样本点间的协方差函数对于所有距离 h 均等于 0，变量的空间相关不存在。

（2）球状模型。其一般公式为：

$$\gamma(h)=\begin{cases}0 & (h=0)\\ c_0+c\left(\dfrac{3h}{2a}-\dfrac{h^3}{2a^3}\right) & (0<h\leqslant a)\\ c_0+c & (h>a)\end{cases} \tag{10.23}$$

式中，c_0 为块金(效应)常数，c 为拱高，c_0+c 为基台值，a 为变程。当 $c_0=0$，$c=1$时，称为标准球状模型。球状模型是空间统计分析中应用最广泛的理论模型，许多区域化变量的理论模型都可以用该模型去拟合。

（3）指数模型。其一般公式为：

$$\gamma(h)=\begin{cases}0 & (h=0)\\ c_0+c(1-\mathrm{e}^{-\frac{h}{a}}) & (h>0)\end{cases} \tag{10.24}$$

式中，c_0 和 c 意义与前相同，但 a 不是变程。当 $h=3a$ 时，$1-\mathrm{e}^{-\frac{h}{a}}=1-\mathrm{e}^{-3}\approx 0.95\approx 1$，即：$\gamma(3a)\approx c_0+c$，从而指数模型的变程 a' 约为 $3a$，当 $c_0=0$，$c=1$ 时，称为标准指数模型。

（4）高斯模型。其一般公式为：

$$\gamma(h)=\begin{cases}0 & (h=0)\\ c_0+c\left(1-\mathrm{e}^{-\frac{h^2}{a^2}}\right) & (h>0)\end{cases} \tag{10.25}$$

式中，c_0 和 c 意义与前相同，a 也不是变程。当 $h=\sqrt{3}a$ 时，$1-\mathrm{e}^{-\frac{h^2}{a^2}}=1-\mathrm{e}^{-3}\approx 0.95\approx 1$,，即 $\gamma(\sqrt{3}a)\approx c_0+c$，因此高斯模型的变程 a' 约为 $3a$。当 $c_0=0$，$c=1$ 时，称为标准高斯函数模型。

（5）幂函数模型。其一般公式为：

$$\gamma(h)=Ah^{\theta}\qquad (0<\theta<2) \tag{10.26}$$

式中，θ 为幂指数。当 θ 变化时，这种模型可以反映在原点附近的各种性状。但是 θ 必须小于 2，若 $\theta\geqslant 2$，则函数 $\gamma(-h)$ 不再是一个条件非负定函数了，也就是说它已经不能成为变差函数了。

(6) 对数模型。其一般公式为：

$$\gamma(h)=A\lg h \tag{10.27}$$

显然，当 $h\to 0$，$\lg h\to -\infty$，这与变差函数的性质 $\gamma(h)\geqslant 0$ 不符。因此，对数模型不能描述点支撑上的区域化变量的结构。

(7) 线性有基台值模型。其一般公式为：

$$\gamma(h)=\begin{cases} c_0 & (h=0) \\ Ah & (0<h\leqslant a) \\ c_0+c & (h>a) \end{cases} \tag{10.28}$$

该模型的变程为 a，基台值为 c_0+c。

(8) 线性无基台值模型。其一般公式为：

$$\gamma(h)=\begin{cases} c_0 & (h=0) \\ Ah & (h>0) \end{cases} \tag{10.29}$$

该模型没有基台值，也没有变程。

在这些变差函数的理论模型中，最常用的是球状模型，球状模型曲线如图 10－1 所示。

变差函数是空间统计学的主要工具，有了变差函数，就可以应用空间统计学的理论和方法对环境空间污染物的空间分布进行研究了。

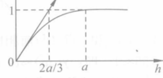


图 10－1 球状模型曲线

4. 变差函数的参数最优估计

变差函数的理论模型主要是曲线模型，将曲线模型经过适当的变换，化为线性模型，然后用最小二乘法原理进行未知参数的估计。表 10.1 是空间统计学中常用的变差函数理论模型经过适当变换后化为的线性模型。对于球状模型、指数模型和高斯模型而言，只讨论 $0<h\leqslant a$ 时的拟合问题。对变换后的变差函数理论模型，除球状模型为二元线性回归模型外，其余均为一元线性回归模型。根据最小二乘法原理对这两类线性回归模型进行参数估计计算。

表 10.1　常用变差函数理论模型的线性变换

变差函数理论模型	变换	变换后的线性模型
球状模型 $\gamma(h)=\begin{cases} 0 & (h=0) \\ c_0+c\left(\dfrac{3h}{2a}-\dfrac{h^3}{2a^3}\right) & (0<h\leqslant a) \\ c_0+c & (h>a) \end{cases}$	$\gamma(h)=y$，$c_0=b_0$ $h=x_1$，$\dfrac{3c}{2a}=b_1$ $h^3=x_2$，$\dfrac{-c}{2a^3}=b_2$	$y=b_0+b_1x_1+b_2x_2$

续表

变差函数理论模型	变换	变换后的线性模型
指数模型 $\gamma(h)=\begin{cases}0 & (h=0)\\ c_0+c(1-e^{-\frac{h}{a}}) & (h>0)\end{cases}$	$\gamma(h)=y$，$e^{-\frac{h}{a}}=x$ $c_0+c=b_0$ $-c=b_1$	$y=b_0+b_1x$
高斯模型 $\gamma(h)=\begin{cases}0 & (h=0)\\ c_0+c\left(1-e^{-\frac{h^2}{a^2}}\right) & (h>0)\end{cases}$	$\gamma(h)=y$，$e^{-\frac{h^2}{a^2}}=x$ $c_0+c=b_0$ $-c=b_1$	$y=b_0+b_1x$

在空间统计学的理论模型中，只有球状模型线性化后成为二元线性回归模型。它共有三个参数 b_0，b_1 和 b_2。如果采用最简单的最小二乘法来作最优参数估计，方法比较方便，但结果得到的变差函数理论模型曲线有时并不十分满意，主要是对实际变差函数曲线中头几个点的重要性认识不够。实际上，变差函数曲线上头几个点(即在原点附近的几个点)的重要性远大于曲线其他点的重要性。不应该把它们与其他实际变差函数曲线上的点平等对待。在原点附近的几个点都在变程范围内，在反映变量的空间自相关性方面极为重要。为了克服这个问题，采用加权回归的方法比较合适，拟合度较高。权重系数主要是采用每一距离上的样本对数 $N(h_i)$。因此，采用加权多项式回归方法进行二元线性回归模型的参数估计。设二元线性回归模型为：

$$y=b_0+b_1x_1+b_2x_2 \tag{10.30}$$

式中，b_0，b_1 和 b_2 为待估参数，加权最小二乘法的参数最优估计公式是：

$$\begin{cases}b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2\\ b_1=(L_{1y}L_{22}-L_{2y}L_{12})/(L_{11}L_{22}-L_{12}L_{21})\\ b_2=(L_{2y}L_{11}-L_{1y}L_{21})/(L_{11}L_{22}-L_{12}L_{21})\end{cases} \tag{10.31}$$

式中：

$$\bar{y}=\sum_{i=1}^{n}N(h_i)y_i/\sum_{i=1}^{n}N(h_i)$$

$$\bar{x}_1=\sum_{i=1}^{n}N(h_i)x_{1i}/\sum_{i=1}^{n}N(h_i)$$

$$\bar{x}_2=\sum_{i=1}^{n}N(h_i)x_{2i}/\sum_{i=1}^{n}N(h_i)$$

$$L_{11}=\sum_{i=1}^{n}N(h_i)(x_{1i}-\bar{x}_1)^2$$

$$L_{22}=\sum_{i=1}^{n}N(h_i)(x_{2i}-\bar{x}_2)^2$$

$$L_{12}=L_{21}=\sum_{i=1}^{n}N(h_i)(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)$$

$$L_{1y}=\sum_{i=1}^{n}N(h_i)(x_{1i}-\bar{x}_1)(y_i-\bar{y})$$

$$L_{2y}=\sum_{i=1}^{n}N(h_i)(x_{2i}-\bar{x}_2)(y_i-\bar{y})$$

$$L_{yy}=\sum_{i=1}^{n}N(h_i)(y_i-\bar{y})^2$$

计算出 b_0，b_1 和 b_2 后还要分三种情况加以讨论。

(1)$b_0>0$，$b_1>0$，$b_2<0$，此时球状模型中三个参数 c_0，c 和 a 分别为：

$$\begin{cases} c_0=b_0 \\ a=\sqrt{\dfrac{-b_1}{3b_2}} \\ c=\dfrac{2b_1}{3}\sqrt{\dfrac{-b_1}{3b_2}} \end{cases} \tag{10.32}$$

这三个参数为最优拟合球状模型时的三个参数。

(2)$b_0<0$，$b_1>0$，$b_2<0$，此时 $b_0<0$，即 $c_0<0$，不符合球状模型的要求，可设 $b_0=0$，这时式(10.31)为 $y=b_1x_1+b_2x_2$，重新根据最小二乘法求出参数 b_1 和 b_2，在 $b_0=0$ 的条件下仍可求出 c_0，c 和 a 三个参数。

(3)$b_0>0$，$b_1>0$，$b_2\geqslant0$，此时应分两种情况，一种是 $b_2=0$，二元线性回归模型(10.30)变为 $y=b_0+b_1x_1$，为一元线性模型，而不是球状模型，可按一元线性回归模型参数估计的方法求解其参数；另一种是 $b_2>0$，这时对原始数据进行调整，增加或删减一些不重要的实际变差函数点的数据，反复多次地调整，直到 $b_2<0$ 时为止，然后代入式(10.12)，求出 c_0，c 和 a 三个参数。

5. 回归模型的检验

通过样本数据建立变差函数理论模型，仅仅进行参数的最优估计是不够的，还必须对回归模型进行显著性检验，这样才能使变差函数理论模型有意义。

(1)用残差平方和或标准误差检验回归模型方程的显著性

实际观测值 y 和理论模型计算出的理论值 $\hat{y}$ 之差，即 $(y-\hat{y})$ 称为残差，残差平方和方程为：

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10.33}$$

其回归估计的标准误差为：

$$S=\sqrt{\frac{Q}{n-2}}=\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{10.34}$$

残差平方和或标准误差愈小，说明实际观测值与回归线愈靠近，拟合的曲线与实际配合愈好；反之，说明配合的理论曲线与实际误差较大。因此，残差平方和或标准误差的值是回归曲线的重要参数。

(2)回归模型的 F 检验

在线性回归条件下，总平方和可以分解为残差平方和与回归平方和两部分，即：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{10.35}$$

这样可构成一个 F 统计量，即：

$$F=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2/(k-1)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-k)} \tag{10.36}$$

式中，k 为回归模型中自变量个数，对于一元线性回归 F 的自由度为 $f=(1,n-2)$。若计算的 F 值大于显著性水平 α(0.05 或 0.01)与自由度 f 的临界值 $F_{\alpha,f}$时，则在显著性水平 α 与自由度 f 时，表明所建立的回归方程与回归直线是显著的，所配合的理论曲线是有意义的。显然，F 值愈大愈显著，回归模型的精度愈高。

(3)回归模型的相关系数和决定系数

回归模型的相关系数 R 的大小，说明自变量 x 和因变量 y 之间线性关系的程度。对于线性回归模型，可以用相关系数 R 的大小来判断回归模型的精度，但是要判断回归模型，尤其是曲线回归模型拟合的好坏，主要是采用决定系数 R^2。决定系数 R^2 是回归平方和占总平方和的百分比。R^2 愈大，该回归模型配合的理论曲线的精度愈高；反之，该回归模型配合的理论曲线精度就愈低，该回归模型的实际意义就不大。那么 R^2 多大，回归模型才有价值呢？还是采用 R^2 的 F 检验。设 F 检验决定系数的统计量为：

$$F=\frac{R^2}{1-R^2}\cdot\frac{n-k}{k-1} \tag{10.37}$$

式中，k 为回归模型中自变量个数，对于一元线性回归 F 的自由度为 $f=(1,n-2)$。若计算的 F 值大于显著性水平 α(0.05 或 0.01)与自由度 f 的临界值 $F_{\alpha,f}$时，R^2 是有意义的，即回归模型配合的理论曲线拟合度较高，可以采用该回归

模型作为理论曲线的模型；反之，该回归模型作为理论曲线模型毫无实际意义。

一般来讲，在空间统计学中变差函数理论模型的最优拟合，通常要选择几个不同的理论模型来比较，从中选出一个最优的理论模型。

10.2.4 普通克立格插值

用于估值的方法种类繁多，常用的有多角形法、剖面法、算术平均值法以及距离平方反比法等，这些估值方法在空间插值应用中有一定的局限性。空间统计学与上述常规方法有着明显的不同。它基于这样一种概念，即用于推断现象的样品相互间不是独立的，它们之间存在着一定的相关关系。这种相关性除了随样品距离变化外，还随样品间的相对方向的变化而变化。它是建立在变差函数理论及结构分析基础上，在有限区域内对区域化变量的取值进行无偏最优估计的一种方法。克立格法是空间统计学的核心。

克立格法(Kriging)也称空间局部估计或空间局部插值，是空间统计学中两大主要方法之一。它是建立在变差函数理论及结构分析基础上，在有限区域内对区域化变量的取值进行无偏最优估计的一种方法。这种方法最早由南非矿业工程师克立格和统计学家西舍尔在 20 世纪 50 年代根据样本空间位置的不同和样本间相关程度的不同，对每个样本赋予一定的权重，进行滑动加权平均，来估计未知样点上样本平均值的一种方法。

克立格法实质上是利用区域化变量的原始数据和变差函数的结构特点，对未采样点的区域化变量的取值进行线性无偏最优估计的一种方法。从数学的角度讲就是一种对空间分布的数据求线性最优无偏内插估计量(best linear unbiased estimator，简写为 BLUE)的一种方法。更具体地讲，它是根据待估样点(或待估块段)有限邻域内若干已测定的样点数据，在认真考虑了样点的形状、大小和空间相互位置关系，它们与待估样点间相互空间位置关系以及变差函数提供的结构信息之后，对该待估样点值进行的一种线性无偏最优估计。

传统的估计方法中常用的多边形法，主要是根据多边形块段内的一个采样资料来估计数值，其缺点是没有考虑周围其他采样点的信息，可说是“一孔之见”；剖面法和三角形法中所利用的每一个采样数据在估值计算中的贡献是一样的，即都是等权的，没有区别不同情况给以不同的权重系数，这就是它们的不足之处；距离反比法(或距离平方反比法)虽然前进了一步，考虑了周围的样品，而且也以各数据用样品到待估块段中心的距离(或距离平方)的倒数为权进行了加权平均，但它们还没有考虑样品彼此之间和样品与待估块段之间的空间几何构形因素的影响，同时也没有考虑到所研究变量的空间分布结构信息(即变差函数)。克立格法

与传统的估计不同，它最大限度地利用了空间取样所提供的各种信息，在估计未知样点数值时，它不仅考虑了落在该样点的数据，而且还考虑了临近样点的数据，不仅考虑了待估样点与临近已知样点的空间位置，还考虑了各临近样点彼此之间的位置关系。除了上述的几个因素外，还利用了已有观测值空间分布的结构特征，使克立格估计比其他传统的估计方法更精确，更符合实际，并且避免系统误差的出现，给出估计误差和精度。这些是克立格法的最大优点。但是，如果变差函数和相关分析的结果表明区域化变量的空间相关性不存在，则空间局部插值的方法不适用。

克立格法是多种多样的，且其本身也在不断发展、完善之中。对各种不同的目的和不同的条件，可以采用各种不同的克立格法，这样可以取得更好的效果。在满足二阶平稳（或本征）假设时可用普通克立格法（ordinary Kriging，简称OK）。在非平稳（或说有漂移存在）现象中，可应用泛克立格法；在计算局部估值时要用到非线性估计量，就可用析取克立格法。此外，当区域化变量服从对数正态分布时，可用对数正态克立格法；对有多个变量的协同区域化现象，可用协克立格法等。其中，最常用的是普通克立格法。

10.2.4.1 一般问题及其解法

设 $Z(x)$ 为区域化变量，满足二阶平稳和本征假设，其数学期望为常数 m，协方差函数 $C(h)$ 和变差函数 $\gamma(h)$ 存在，即：

$$\begin{aligned} &E[Z(x)]=m \\ &C(h)=E[Z(x)Z(x+h)]-m^2 \\ &\gamma(h)=\frac{1}{2}E[Z(x)-Z(x+h)]^2 \end{aligned} \tag{10.38}$$

对中心位于 x_0 的块段 V 的平均值 $Z_V(x_0)$ 以

$$Z_V(x_0)=\frac{1}{V}\int_V Z(x)\mathrm{d}x \tag{10.39}$$

进行估值。在待估块段 V 的邻域内，$Z_i(i=1, 2, 3, \cdots, n)$ 是一组离散的信息样品数据，它们是定义在点承载 $x_i(i=1, 2, \cdots, n)$ 上的；或是确定在以 x_i 点为中心的承载 v_i 上的平均值 $Z_{v_i}(x_i)$（简记为 Z_i）。且这 n 个承载 $v_i(i=1, 2, \cdots, n)$ 既不同于 V，又各不相同（图 10-2）。

进行估计所使用的线性估计量为：

$$Z_V^* = \sum_{i=1}^{n}\lambda_i Z_i \tag{10.40}$$

它是 n 个数值的线性组合。

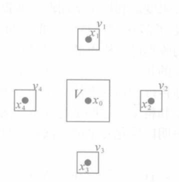


图 10－2　$n=4$ 时信息样点和待估块段承载图(或估计构形图)

克立格估值的原则，就是在保证估计量 Z_V^* 是无偏的，且估计方差最小的前提下，求出 n 个权系数 λ_i。

10.2.4.2　普通克立格法

当区域化变量 $Z(x)$ 的数学期望 $E[Z(x)]=m$ 为未知常数时，实际上在研究之前也常常如此，这时的估计采用普通克立格法。若要使 Z_V^* 为 Z_V 的无偏估计量，即要求：

$$E(Z_V^*-Z_V)=0 \tag{10.41}$$

因为　$E(Z_V)=\frac{1}{V}\int_V E[Z(x)]\mathrm{d}x=m$

又因为　$E(Z_V^*)=E\left(\sum_{i=1}^{n}\lambda_i Z_i\right)=\sum_{i=1}^{n}\lambda_i E(Z_i)=m\sum_{i=1}^{n}\lambda_i$

故得无偏性条件：

$$\sum_{i=1}^{n}\lambda_i=1 \tag{10.42}$$

在满足无偏性条件下，估计方差 σ_E^2 为：

$$\begin{aligned}\sigma_E^2&=E(Z_V-Z_V^*)^2=E\left[Z_V-\sum_{i=1}^{n}\lambda_i Z(x_i)\right]^2\\&=\overline{C}(V,V)+\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\overline{C}(x_i,x_j)-2\sum_{i=1}^{n}\lambda_i\overline{C}(x_i,V)\end{aligned} \tag{10.43}$$

要使估计方差 σ_E^2 为最小，根据拉格朗日原理，令：

$$F=\sigma_E^2-2\mu\left(\sum_{i=1}^{n}\lambda_i-1\right) \tag{10.44}$$

这里，F 是 n 个权系数 λ_i 和 μ 的 $(n+1)$ 元函数，-2μ 是拉格朗日乘数。求出 F 对 $\lambda_i(i=1, 2, \cdots, n)$ 以及 F 对 μ 的偏导数，并令其为零，便得到下列方程组：

$$\begin{cases} \dfrac{\partial F}{\partial \lambda_i} = -2\overline{C}(x_i, V) + 2\sum_{j=1}^{n}\lambda_j C(x_i, x_j) - 2\mu = 0 \\ \dfrac{\partial F}{\partial \mu} = -2\left(\sum_{i=1}^{n}\lambda_i - 1\right) = 0 \end{cases} \tag{10.45}$$

整理得：

$$\begin{cases} \sum_{j=1}^{n}\lambda_j C(x_i, x_j) - \mu = \overline{C}(x_i, V) \\ \sum_{i=1}^{n}\lambda_i = 1 \end{cases} \tag{10.46}$$

这 $(n+1)$ 个方程的方程组，称为普通克立格方程组。

普通克立格方差计算公式为：

$$\sigma_K^2 = \overline{C}(V, V) - \sum_{i=1}^{n}\lambda_i\overline{C}(x_i, V) + \mu \tag{10.47}$$

用变差函数表示为：

$$\begin{cases} \sum_{j=1}^{n}\lambda_j \gamma(x_i, x_j) + \mu = \overline{\gamma}(x_i, V) \\ \sum_{i=1}^{n}\lambda_i = 1 \end{cases} \tag{10.48}$$

$$\sigma_K^2 = \sum_{i=1}^{n}\lambda_i\overline{\gamma}(x_i, V) - \overline{\gamma}(V, V) + \mu \tag{10.49}$$

以上样品的承载是点承载的情况，若样品的承载是以 x_i 为中心，其体积为 v_i 的承载时，将公式中的协方差 $C(x_i, x_j)$ 变为样品域之间的平均协方差 $\overline{C}(v_i, v_j)$，相应的公式为：

$$\begin{cases} \sum_{j=1}^{n}\lambda_j \overline{C}(v_i, v_j) - \mu = \overline{C}(v_i, V) \\ \sum_{i=1}^{n}\lambda_i = 1 \end{cases} \quad (i = 1, 2, \cdots, n) \tag{10.50}$$

$$\sigma_K^2 = \overline{C}(V, V) - \sum_{i=1}^{n}\lambda_i\overline{C}(v_i, V) + \mu \tag{10.51}$$

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \bar{\gamma}(v_i, v_j) + \mu = \bar{\gamma}(v_i, V) \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \quad (i = 1, 2, \cdots, n) \tag{10.52}$$

$$\sigma_K^2 = \sum_{i=1}^{n} \lambda_i \bar{\gamma}(v_i, V) - \bar{\gamma}(V, V) + \mu \tag{10.53}$$

上述过程也可用矩阵形式表示，令：

$$\boldsymbol{K} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}, \quad \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} c(x_1, x) \\ c(x_2, x) \\ \vdots \\ c(x_n, x) \\ 1 \end{pmatrix}$$

则普通克立格方程组为：

$$\boldsymbol{K\lambda} = \boldsymbol{D} \tag{10.54}$$

解方程组(10.54)，可得：

$$\boldsymbol{\lambda} = \boldsymbol{K}^{-1}\boldsymbol{D} \tag{10.55}$$

其估计方差为：

$$\sigma_K^2 = C(x, x) - \boldsymbol{\lambda}'\boldsymbol{D} \tag{10.56}$$

在以上的介绍中，区域化变量 $Z(x)$ 的数学期望 $E[Z(x)] = m$ 可以是已知或未知的。如果 m 是已知常数，称为简单克立格法；如果 m 是未知常数，称为普通克立格法。不管是哪一种方法，均可根据以上方法计算权重系数和克立格估计量。

10.2.5 环境应用

例 10.1 变差函数计算实例(徐建华，2006)

假设某地区降水量 $Z(x)$(单位：mm)是二维区域化随机变量，满足二阶平稳假设，其观测值的空间正方形网格数据如图 10－3 所示(点与点之间的距离为 $h=1$ km)。试计算其南北方向及西北和东南方向的变差函数。

从图 10－3 可以看出，空间上有些点，由于某种原因没有采集到。如果没有缺失值，可直接对正方形网格数据结构计算变差函数；在有缺失值的情况下，也可以计算变差函数。只要“跳过”缺失点位置即可(图 10－4)。

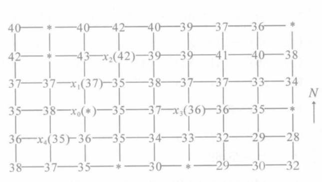


图 10－3　空间正方形网格数据(点间距 h=1 km)

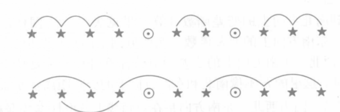


图 10－4　缺失值情况下样本数对的组成和计算过程(⊙为缺失值)

首先计算南北方向上的变差函数值，由变差函数的计算公式(10.21)可得：

$$\gamma(1)=\frac{1}{2\times36}[(40-42)^2+(42-37)^2+(37-35)^2+(35-36)^2+(36-38)^2+(37-38)^2+(38-35)^2+(35-37)^2+(40-43)^2+(43-37)^2+(36-35)^2+(42-42)^2+(42-35)^2+(35-35)^2+(35-35)^2+(40-39)^2+(39-38)^2+(38-37)^2+(37-34)^2+(34-30)^2+(39-39)^2+(39-37)^2+(37-36)^2+(36-33)^2+(37-41)^2+(41-37)^2+(37-36)^2+(36-32)^2+(32-29)^2+(36-40)^2+(40-33)^2+(33-35)^2+(35-29)^2+(29-30)^2+(38-34)^2+(28-32)^2]$$

$$=385/72=5.35$$

$$\gamma(2)=\frac{1}{2\times27}[(40-37)^2+(42-35)^2+(37-36)^2+(35-38)^2+(37-35)^2+(38-37)^2+(40-37)^2+(37-36)^2+(42-35)^2+(42-35)^2+(35-35)^2+(40-38)^2+(39-37)^2+(38-34)^2+(37-30)^2+(39-37)^2+(39-36)^2+(37-33)^2+(37-37)^2+(41-36)^2+(37-32)^2+(36-29)^2+(36-33)^2+(40-35)^2+(33-29)^2+(35-30)^2+(34-28)^2]=493/54=9.13$$

$$\gamma(3)=\frac{1}{2\times 21}(40-35)^2+(42-36)^2+(37-38)^2+(37-37)^2+(43-36)^2+(37-35)^2+(42-35)^2+(42-35)^2+(40-37)^2+(39-34)^2+(38-30)^2+(39-36)^2+(39-33)^2+(37-36)^2+(41-32)^2+(37-29)^2+(36-35)^2+(40-29)^2+(33-30)^2+(38-28)^2+(34-32)^2]=737/42=17.55$$

$$\gamma(4)=\frac{1}{2\times 13}[(40-36)^2+(42-38)^2+(40-36)^2+(43-35)^2+(42-35)^2+(40-34)^2+(39-30)^2+(39-33)^2+(37-32)^2+(41-29)^2+(36-29)^2+(40-30)^2+(38-32)^2]=668/26=25.69$$

$$\gamma(5)=\frac{1}{2\times 5}[(40-38)^2+(40-35)^2+(40-30)^2+(37-29)^2+(36-30)^2]=229/10=22.90$$

最后，得到南北方向上的变差函数计算结果见表 10.2。同样，可以计算东西方向和西北一东南方向上的变差函数。东西方向上的计算与南北方向相同(这里不再赘述)。西北一东南方向上的变差函数的计算过程，主要是找出分隔距离 h' 和样本数据对。这里的 h' 不像南北和东西方向上的分隔距离 h 是整数，而是对角线上的距离 $\sqrt{2}h$。因为西北一东南方向是在对角线上选取样本数据对 $N(h)$，对正方形网格数据每一分隔距离都要乘以 $\sqrt{2}$，变差函数的计算方法与前面均相同。譬如 $\gamma(5\sqrt{2})$ 的计算过程为：

$$\gamma(5\sqrt{2})\approx\gamma(7.07)=\frac{1}{2\times 2}\left[(42-32)^2+(40-30)^2\right]=200/4=50.00$$

采用同样方法计算获得的西北一东南方向上变差函数的其他计算结果(表 10.2)。

表 10.2　南北、西北一东南方向上的变差函数计算结果

方向	南北					方向	西北一东南				
h	1	2	3	4	5	h	1.41	2.82	4.24	5.65	7.07
$N(h)$	36	27	21	13	5	$N(h)$	32	21	13	8	2
$\gamma(h)$	5.35	9.13	17.55	25.69	22.90	$\gamma(h)$	7.06	12.95	30.85	58.13	50.00

从上面的介绍和讨论，我们知道，球状变差函数的一般形式为：

$$\gamma(h)=\begin{cases}0 & (h=0)\\ c_0+c\left(\dfrac{3h}{2a}-\dfrac{h^3}{2a^3}\right) & (0<h\leqslant a)\\ c_0+c & (h>a)\end{cases}$$

当 $0<h\leqslant a$ 时，有：

$$\gamma(h)=c_0+\left(\frac{3c}{2a}\right)h-\left(\frac{c}{2a^3}\right)h^3$$

如果记 $y=\gamma(h)$，$b_0=c_0$，$b_1=\frac{3c}{2a}$，$b_2=-\frac{c}{2a^3}$，$x_1=h$，$x_2=h^3$，则可以得到线性模型：

$$y=b_0+b_1x_1+b_2x_2 \tag{10.57}$$

根据表 10.2 中的数据，对式(10.58)进行最小二乘拟合，得到：

$$y=2.048+1.731x_1-0.007\ 92x_2 \tag{10.58}$$

进一步计算可知，式(10.58)的显著性检验参数 $F=114.054$，$R^2=0.962$，可见模型的拟合效果是很好的。

比较式(10.57)与式(10.58)，并做简单计算可知：$c_0=2.048$，$c=1.154$，$a=8.353$，所以，球状变差函数模型为：

$$\gamma^*(h)=\begin{cases}0 & (h=0)\\ 2.048+1.154\left(\dfrac{3h}{2\times 8.535}-\dfrac{h^3}{2\times 8.535^3}\right) & (0<h\leqslant 8.535)\\ 3.202 & (h>8.535)\end{cases} \tag{10.59}$$

在实际分析问题中，变差函数模型的拟合计算，一般需要借助于有关软件来完成。

例 10.2　克立格估计实例(徐建华，2006)

以图 10－3 为例，4 个观测点 x_1，x_2，x_3，x_4 的观测值分别为 $Z(x_1)=37$，$Z(x_2)=42$，$Z(x_3)=36$，$Z(x_4)=35$，如果假设降水量的变差函数是各向同性(变差函数在各个方向的变化都相同)的二维球状模型，其具体形式为式(10.29)。现在，我们用普通克立格法估计观测点 x_0 的降水量值 $Z(x_0)$。

根据普通克立格法的基本原理，我们知道，$Z(x_0)$ 估计的基本公式应该是：

$$Z^*(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i)\quad (i=1,\ 2,\ 3,\ 4)$$

根据式(10.55)，可知：

$$\begin{pmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\\ \lambda_4\\ \mu\end{pmatrix}=\begin{pmatrix}c_{11} & c_{12} & c_{13} & c_{14} & 1\\ c_{21} & c_{22} & c_{23} & c_{24} & 1\\ c_{31} & c_{32} & c_{33} & c_{34} & 1\\ c_{41} & c_{42} & c_{43} & c_{44} & 1\\ 1 & 1 & 1 & 1 & 0\end{pmatrix}^{-1}\begin{pmatrix}c_{01}\\ c_{02}\\ c_{03}\\ c_{04}\\ 1\end{pmatrix} \tag{10.60}$$

根据协方差与变差函数的关系以及式(10.59)，可得协方差函数：

$$c^*(h)=\begin{cases}3.202 & (h=0)\\ 1.154\left[1-\left(\dfrac{3h}{2\times 8.535}-\dfrac{h^3}{2\times 8.535^3}\right)\right] & (0<h\leqslant 8.535)\\ 0 & (h>8.535)\end{cases}$$

当 $i=j$ 时，$c_{11}=c_{22}=c_{33}=c_{44}=c(0)=c_0+c=2.048+1.154=3.202$

根据克立格矩阵的对称性，当 $i\neq j$ 时，$c_{ij}=c(|x_i-x_j|)=3.202-\gamma|x_i-x_j|$，由此计算可得：

$$c_{12}=c_{21}=c_{04}=3.202-\gamma(\sqrt{1^2+1^2})=3.202-\gamma(\sqrt{2})$$

$$=3.202-\left[2.048+1.154\left(\frac{3}{2}\times\frac{\sqrt{2}}{8.535}-\frac{1}{2}\times\frac{(\sqrt{2})^3}{8.535^3}\right)\right]=0.870$$

$$c_{14}=c_{41}=c_{02}=3.202-\gamma(\sqrt{2^2+1^2})=0.711$$

$$c_{23}=c_{32}=3.202-\gamma(\sqrt{2^2+2^2})=0.601$$

$$c_{34}=c_{43}=3.202-\gamma(\sqrt{4^2+1^2})=0.383$$

$$c_{01}=3.202-\gamma(\sqrt{1^2})=0.952$$

$$c_{03}=3.202-\gamma(\sqrt{3^2})=0.571$$

将以上计算结果代入克立格方程组(10.54)，得：

$$\begin{pmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\\ \lambda_4\\ \mu\end{pmatrix}=\begin{bmatrix}3.202 & 0.870 & 0.542 & 0.711 & 1.000\\ 0.870 & 3.202 & 0.601 & 0.466 & 1.000\\ 0.542 & 0.601 & 3.202 & 0.383 & 1.000\\ 0.711 & 0.466 & 0.383 & 3.202 & 1.000\\ 1.000 & 1.000 & 1.000 & 1.000 & 0.000\end{bmatrix}^{-1}\begin{bmatrix}0.952\\ 0.711\\ 0.571\\ 0.870\\ 1.000\end{bmatrix}=\begin{bmatrix}0.287\\ 0.210\\ 0.202\\ 0.301\\ -0.473\end{bmatrix}$$

即克立格权重系数分别为：$\lambda_1=0.287$，$\lambda_2=0.210$，$\lambda_3=0.202$，$\lambda_4=0.301$，$\mu=-0.473$，所以 x_0 点的降水量的克立格估计值为：

$$Z_0^*=0.287Z(x_1)+0.210Z(x_2)+0.202Z(x_3)-0.301Z(x_4)$$
$$=0.287\times 37+0.210\times 42+0.202\times 36-0.301\times 35=37.250(\text{mm})$$

克立格估计方差为：

$$\sigma_K^2=c(x_0,x_0)-\sum_{i=1}^{4}\lambda_i c(x_i,x_0)+\mu$$
$$=3.202-(0.287\times 0.952+0.210\times 0.711+0.202\times 0.571+0.301\times 0.870)+0.473$$
$$=2.875(\text{mm})$$

在实际分析问题中，克立格插值计算量往往较大，需要借助于有关软件来完成，目前，在 ArcGIS8.0 以上版本的 geostatistical analyst 模块中，借助于

geostatistical 的 wizard 向导，不但可以完成普通克立格的插值计算，还可以实现泛克立格、指示克立格、析取克立格、协同克立格等方法的插值计算过程。

例 10.3　软件计算实例

(1)数据采集

以太湖水质监测为例，对水质参数进行环境空间统计分析。在采样监测中，考虑到湖岸线和湖中岛的影响，在太湖中均匀布置 12 行 10 列共 75 个采样点，南北向采样点间距为 5.585 km，东西向采样点间距为 6.079 km。采样和样品的分析根据国家环境有关规范进行。在监测中主要考虑了叶绿素 a(chlorophyll a，简称 chl-a)、总悬浮物(suspended sediment，简称 SS)、透明度(secchi depth，简称 SD)三种水质参数。



图 10－5　采样点分布图

(2)异常值的识别与处理——影响系数法

影响系数法是在研究区域化变量变异程度基础上，对可能出现的异常值的影响系数 k 人为赋值，以适当地抑制其影响程度的一种异常值识别与处理方法。该法对样品组观测值需进行多次识别后方能识别出所有异常值。其具体步骤如下：

首先，针对湖泊水质参数区域化变量的 n 个观测值，分别计算其均值 M 和去掉可疑值的 $n-1$ 个观测值的均值 m。以太湖采样获得的水质参数——总悬浮物(SS)的部分采样值为例，其原始数据列及 M 和 m 的计算值见表 10.3。

表 10.3　　影响系数法计算结果

采样点	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}	Z_{16}	Z_{17}	Z_{18}	Z_{19}	Z_{20}
SS	31	28	44	28	25	24	18	17	30	27
M	27.55	27.55	27.55	27.55	27.55	27.55	27.55	27.55	27.55	27.55
m	27.368 4	27.526 3	26.684 2	27.526 3	27.684 2	27.736 8	28.052 6	28.105 2	27.421 1	27.579 0
M/m	1.006 6	1.000 9	1.032 5	1.000 9	0.995 2	0.993 3	0.982 1	0.980 2	1.004 7	0.999 0
采样点	Z_{21}	Z_{22}	Z_{23}	Z_{24}	Z_{25}	Z_{26}	Z_{27}	Z_{28}	Z_{29}	Z_{30}
SS	38	30	22	26	12	24	26	20	38	43
M	27.55	27.55	27.55	27.55	27.55	27.55	27.55	27.55	27.55	27.55
m	27.000 0	27.421 1	27.842 1	27.631 6	28.368 4	27.736 8	27.631 6	27.947 4	27.000 0	26.736 8
M/m	1.020 4	1.004 7	0.989 5	0.997 1	0.971 2	0.993 3	0.997 1	0.985 8	1.020 4	1.030 4

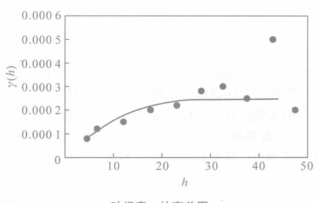
其次，根据观测值的变异性，对影响系数 k 人为赋值，当 $M/m \leqslant k+1$ 时，可疑值不为异常值，否则该值被确定为异常值。以表 10.3 数据列为例，当 $k=0.05$ 时有 $M/m \leqslant 1.05$，在所有的检验数据中，没有一个的结果大于 1.05，故判定为没有异常值，即所检验的数据对全部样品值的影响没有一个超过了 5%。

然后，如果有异常值，用异常值下限值 $GL=M[(nk+1)/(k+1)]$ 代替异常值。本例 $k=0.1$ 时，$GL=75.14$；$k=0.05$ 时，$GL=52.48$。

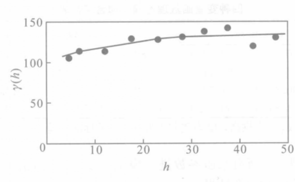
如果有异常值，重复上述步骤，直到再也识别不出新的异常值为止。

(3) 实验变差函数

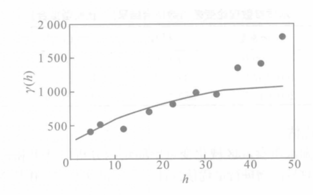
根据实验变差函数计算公式，计算出实验变差函数值，作出变差图(图 10-6)。



叶绿素 a 的变差图



总悬浮物的变差图



透明度的变差图

图 10-6 各水质参数的结构分析变差图

(4) 理论变差函数拟合

理论模型的最优拟合最重要的是对模型中的参数进行最优估计。变差函数的理论模型主要是曲线模型，将曲线模型经过适当的变换，化为线性模型，然后用最小二乘法原理进行未知参数的估计。通过各个模型之间的比较，可以发现有些模型相对于其他几个模型来说，更符合变差图中实际的点，必须通过它们之间的比较，选择出一个最优的理论模型。在模型间的比较时，选用了残差平方和(Q)、标准误差(S)和决定系数(R^2)等参数。分别应用四种模型对太湖水质参数总悬浮物进行模拟，计算结果见表 10.4。从表 10.4 可以看出，在四个理论模型中，无论是残差平方和(Q)还是标准误差(S)都是球状模型最小(分别是375.624 1和6.852 2)，而球状模型的决定系数(R^2)又是最大的(0.747 0)，尽管它们之间的块金常数和基台值都基本相同，选择球状模型作为太湖水质参数总悬浮物的变差函数理论模型是比较合适的。

表 10.4 四种变差函数理论模型拟合参数表

理论模型	c_0	c	a	Q	S	R^2
线性有基台值模型	100.164 0	50.028 0	44.893 0	1 301.251 9	12.753 7	0.655 2
球状模型	100.107 9	35.114 0	33.829 5	375.624 1	6.852 2	0.747 0
指数模型	90.483 0	57.528 0	51.325 0	2 359.848 9	17.175 0	0.542 3
高斯模型	104.815 0	50.372 0	52.013 0	1 453.124 9	13.477 4	0.516 8

注：c_0 为块金常数，c 为拱高，Q 为残差平方和，S 为标准误差，R^2 为决定系数。

应用球状模型，对所有水质参数进行模拟，表 10.5 给出了各个水质参数的球状模型理论变差函数拟合结果。

表 10.5 水质参数理论变差函数拟合结果(球状模型拟合)

参数	块金常数	拱高	基台值	变程/km
chl-a	0.000 02	0.000 22	0.000 24	29.152 4
SS	100.107 9	35.114 0	135.221 9	33.829 5
SD	200.013 5	850.201 9	1 050.215 4	35.171 2

（5）空间最优估计

应用前面太湖水质参数区域化变量空间结构分析的结果和普通克立格方程组，对水质参数总悬浮物进行最优估计计算，并以其他方法作比较，估计结果见表 10.6。

表 10.6 克立格内插与传统估计方法结果比较

采样点	实际值	线性内插法		距离平方反比法		克立格法	
		估计值	估计误差	估计值	估计误差	估计值	估计误差
6	26	31.50	−5.50	31.54	−5.54	27.96	−1.96
9	26	29.25	−3.25	29.36	−3.36	26.66	−0.66
14	25	26.50	−1.50	26.46	−1.46	26.15	−1.15
15	24	23.75	0.25	23.56	0.44	24.09	−0.09
19	27	29.75	−2.75	30.11	−3.11	29.66	−2.66
26	26	23.00	3.00	22.92	3.08	23.00	3.00
32	30	30.00	0.00	30.34	−0.34	29.63	0.37
33	26	25.75	0.25	25.90	0.10	26.30	−0.30
34	25	25.50	−0.50	25.29	−0.29	26.53	−1.53
45	30	30.50	−0.50	31.18	−1.18	29.67	0.33
估计误差均值		−1.05		−1.17		−0.47	
估计误差方差		4.92		5.15		2.22	

对比各种空间估值方法的结果可以看出，克立格法的估计误差均值和估计误差方差两项指标均最小，这就表明了克立格法在内陆湖泊的实际应用中确比传统空间估计方法有着更强的有效性、最优性和无偏性。克立格法既考虑了内陆湖泊水质参数空间变化的随机性又考虑了变化的结构性(相关性)，是一种用统计的方法揭示变量空间结构性的数学方法。这一本质特征决定了克立格法用于内陆湖泊水质参数空间估计时，具有其他传统方法所不具有的许多优越性质。比如，克立格法能够给出各空间预测点的估计精度，无须事先知道该点的实测值。而传统方法则无法给出估计精度，一般只能用不同方法的计算结果加以比较，当然更谈不上有一种衡量估计精度的标准和方法了。

(6) 太湖水质评价

根据前面太湖水质参数区域化变量空间结构分析的结果，应用克立格法对太湖整个水域进行水质参数空间估计计算。所有的运算都是在计算机上进行的，估计计算的软件是 SURFER 软件。太湖中总悬浮物的评价结果见图 10.7。

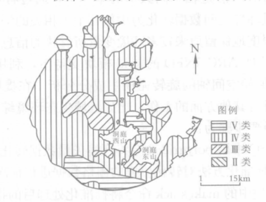


图 10－7　太湖中总悬浮物的评价结果

10.3　环境空间主成分分析

主成分分析就是设法将原来众多的具有一定相关性的指标，重新组合成一组新的相互无关的综合指标来代替原来的指标，而保持其原指标所提供的大量信息(Johnson 等，1998)。主成分分析的基本原理是：将 N 个相关变量 X_i 线性组合成 M 个独立变量 $Y_j(M<N)$，Y_j 中保持了 X_i 中大部分信息，于是 N 个相关变量 X_i 就缩减成 M 个独立变量 Y_j，Y_j 就是通常所说的主成分。

主成分分析需经过以下主要步骤：

(1)原始数据标准化处理。为克服各参数指标量纲的不一致，常用正规化等

处理方法对数据作相应变换；

参评因子的一般标准化量化公式为：

$$Q_i=\frac{X_i-X_{\min}}{X_{\max}-X_{\min}}\times 10 \qquad (i=1,\ 2,\ \cdots,\ N) \tag{10.61}$$

其中，Q_i 为某参评因子的第 i 级的分级标准化值，X_i 为某参评因子的第 i 级编码值，$X_{\min}$ 为参评因子的最小编码值，$X_{\max}$ 为参评因子的最大编码值。

(2)建立 N 个变量的相关系数矩阵 $\boldsymbol{R}$；

(3)计算相关系数矩阵 $\boldsymbol{R}$ 的特征值和相应的单位特征向量；

(4)将特征向量作线性组合，输出 m 个主成分。

10.3.1 空间主成分分析步骤

环境空间主成分分析法是在空间数据的基础上，通过将原始空间坐标轴旋转，将相关的多变量环境空间数据转化为少数几个不相关的综合指标，实现用较少的综合指标最大限度地保留原来较多环境变量所反映的信息。空间主成分分析是在地理信息系统软件 ARC/INFO 的 grid 模块支持下，利用该模块中的 princomp 函数，通过对原始空间轴的旋转完成主成分分析。在提取出来的空间主成分的基础上，可以进行其他方面的工作，比如区域生态环境综合评价、区域生态脆弱性评价等。其重要步骤为：

(1)在 ARC/INFO 中用 polygrid 命令将环境矢量数据转化为栅格数据；

(2)按照一定的标准化方法对转化生成的栅格数据进行标准化处理；

(3)利用 grid 模块中的 makestack 命令将标准化处理后的指标 X_i 图转化为一个综合图；

(4)利用 grid 模块中的 princomp 函数，对综合图进行主成分转换，根据所转换的空间主成分特征向量，利用公式：

$$\alpha_i=\lambda_i/\sum_{i=1}^{m}\lambda_i \qquad (i=1,\ 2,\ \cdots,\ m) \tag{10.62}$$

计算得到各主成分的贡献率，再根据主成分累计贡献率大小，来确定主成分数；

(5)在环境综合评价中，综合评价指数定义为 M 个主成分的加权和，而权重用每个主成分相对应的贡献率来表示，即：

$$E=a_1Y_1+a_2Y_2+\cdots+a_jY_j \qquad (j=1,\ 2,\ \cdots,\ M) \tag{10.63}$$

其中，E 为环境综合评价指数；Y_j 为第 j 个主成分；a_j 为第 j 主成分对应的贡献率。

10.3.2 环境应用

区域生态环境质量是区域经济社会可持续发展的核心和基础，研究流域的区域生态环境质量及其演变，有助于制订和规划流域经济发展计划。区域生态环境是人与自然、环境交互作用的集中体现，受到自然因素和人文因素的共同影响。自然因素和人文因素的各个因子之间相互作用和相互影响，以不同的方式和程度影响着区域生态环境质量的状况。因此，在评价区域生态环境时要充分考虑各种因子的综合作用。

本例题以三峡库区大宁河流域为对象，在遥感和 GIS 基础上应用空间主成分分析方法，综合评价了大宁河流域 1990 年和 2000 年的生态环境质量，并分析了该流域生态环境在这 10 年中的历史演变。

● **研究区概况**

大宁河又名盐溪、昌江，发源于重庆巫溪县境内，于巫山县城以东注入长江，全长202 km，流域面积达4 415.84 km^2，地处长江三峡库区之中。本区地处亚热带湿润区，多年平均降水量1 000 mm左右，年均温 19.8℃。在地貌上属于四川盆地东部边缘山地，地势南北高而中间低。

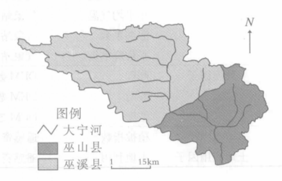


图 10－8 大宁河流域及其行政区域示意图

近几年，由于河流水体污染的原因，大宁河经常出现“水华”现象，对三峡库区造成了一定的影响。保持良好的生态环境不仅是发展流域经济、实现可持续发展的重要基础，也关系到三峡库区生态环境安全，尤其对保障三峡工程的长期有效运行具有重要作用。然而，由于长期以来对环境资源不合理的开发利用，大

宁河流域生态环境已经非常脆弱。因此，大宁河流域的生态环境保护问题受到了普遍重视。

● 评价指标体系和数据获取

影响生态环境质量的因子是多方面的，有自然因素也有人为因素，是典型的自然一经济一社会复合系统。因此，为了有效地综合评价区域生态环境，在制定评价指标体系中要同时考虑自然因素和人为因素两方面。根据以往的研究（Gessler 等，1995；Wilson 等，1996；Bellmann，2000；黄裕婕等，2000；王思远等，2002，2004；左伟，2004），在分析大宁河流域生态环境、地理特征以及空间尺度等特点以及数据可获取性的基础上，从气候、水文、土壤、土地利用、地形地貌等方面选择本次评价的指标。从气候因子、地形因子、植被因子和土地利用因子 4 类影响因子中选取了 10 个评价指标：0℃以上积温、10℃以上积温、年平均气温、湿润系数、平均降水量、高程、坡向、坡度、植被指数、土地利用。指标体系及数据来源见表 10.7。

表 10.7　生态环境综合评价指标体系和数据来源

一级指标	二级指标	三级指标	数据获取
自然因素	气候因子	>0℃积温	气象站点实测资料
		>10℃积温	气象站点实测资料
		年平均气温	气象站点实测资料
		湿润系数	气象站点实测资料计算
		平均降水量	气象站点实测资料
	地形因子	高程	DEM 数据
		坡向	DEM 数据计算
		坡度	DEM 数据计算
人为因素	植被因子	植被指数	遥感资料解译
	土地利用因子	土地利用	遥感资料解译

● 数据栅格化

在对参评因子进行栅格化处理过程中，根据大宁河流域的实际情况，以 100 m×100 m的栅格大小为评价单元，这样生态环境综合评价结果不仅能够反映生态环境质量的高低，而且可以更好地反映生态环境的区域差异。

● 标准化处理

为定量评价大宁河流域生态环境质量，需提取空间数据库中的各图层的专题数据，利用 GIS 软件提供的分析工具，进行主成分分析以及生态环境综合评价。由于各种专题数据性质不同，量纲各异，直接用它们进行评价是困难的。因此在分析和评价之前，需按照一定的标准对参评因子进行标准化处理。根据公式（10.61）进行标准化处理，任何参评因子都被标准化为 1 到 10 之间，消除了量纲的影响，增强了评价结果的可信度。

● 空间主成分分析

在地理信息系统软件 ARC/INFO 的 grid 模块支持下，应用主成分分析功能进行各综合指数的计算。首先将 0℃以上积温、10℃以上积温、年平均气温进行主成分分析生成热量综合指数，将平均降水量、湿润系数进行主成分分析生成水分综合指数，将坡度、坡向、高程进行主成分分析生成地形地貌综合指数，将土地利用、植被指数进行主成分分析生成土地覆被综合指数，然后将热量综合指数和水分综合指数进行主成分分析生成气候综合指数，最后根据气候综合指数、土地覆盖指数和地形地貌指数进行主成分综合评价，计算出大宁河流域的生态环境综合指数。1990 年和 2000 年主成分分析结果见表 10.8。

表 10.8　各主成分的特征值、贡献率和累计贡献率

指标	主成分	1990 年			2000 年		
		特征值	贡献率	累计贡献率	特征值	贡献率	累计贡献率
热量综合	SPCA1	5.679	0.997	0.997	6.225	0.997	0.997
	SPCA2	0.014	0.002	0.999	0.012	0.002	0.999
	SPCA3	0.006	0.001	1.000	0.006	0.001	1.000
水分综合	SPCA1	3.552	0.970	0.970	3.856	0.997	0.997
	SPCA2	0.109	0.030	1.000	0.012	0.003	1.000
地形地貌综合	SPCA1	6.561	0.533	0.533	6.561	0.533	0.533
	SPCA2	3.330	0.271	0.804	3.330	0.271	0.804
	SPCA3	2.415	0.196	1.000	2.415	0.196	1.000
气候综合	SPCA1	5.346	0.688	0.688	5.683	0.700	0.700
	SPCA2	2.421	0.312	1.000	2.436	0.300	1.000
土地覆被综合	SPCA1	8.520	0.970	0.970	7.926	0.756	0.756
	SPCA2	0.267	0.030	1.000	2.553	0.244	1.000
生态环境综合	SPCA1	6.207	0.620	0.620	5.845	0.729	0.729
	SPCA2	2.458	0.246	0.866	1.654	0.206	0.935
	SPCA3	1.340	0.134	1.000	0.522	0.065	1.000

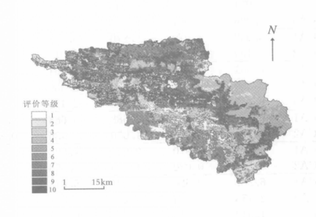
从表 10.8 可以看出，在最终生成的生态环境综合指数中，1990 年和 2000 年的前两个主成分累计贡献率分别为 86.6%和 93.5%，基本上保留了原来变量所反映的主要信息，可信度较高。

● 评价结果及其空间分布

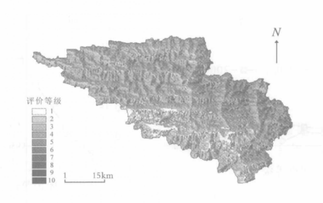
在地理信息系统 ArcGIS 支持下，根据表 10.8 和式(10.63)，可以得到大宁河流域生态环境综合评价等级图(图 10－9)和评价结果统计表(表 10.9)。根据生态环境综合指数大小，将大宁河流域划分为 10 级不同的生态环境质量区。

表 10.9　　大宁河流域生态环境综合评价结果

评价等级	1990 年		2000 年	
	面积/km^2	面积比例	面积/km^2	面积比例
1	290.881	0.066	160.975	0.036
2	254.532	0.058	402.793	0.091
3	397.069	0.090	583.945	0.132
4	459.954	0.104	648.656	0.147
5	146.955	0.033	768.248	0.174
6	212.926	0.048	718.238	0.163
7	627.912	0.142	655.804	0.149
8	61.588	0.014	374.530	0.085
9	1 743.096	0.395	94.957	0.022
10	220.929	0.050	7.697	0.002
合计	4 415.843	1.000	4 415.843	1.000



1990 年



2000 年

图 10－9　大宁河流域生态环境综合评价等级图

从图 10－9 和表 10.9 可以看出，1990 年大宁河流域生态环境综合评价结果在各个等级所占比例上以及空间分布上有很大差异。等级 9 所占比例最大，约 40%，主要分布在上游的巫溪县境内。相比之下，其他等级各自所占比例都很小，分布不具有显著特征。2000 年大宁河流域生态环境综合评价结果在各个等级所占比例上以及空间分布上也有很大差异。没有一个等级所占的比例非常突出，比例相对较大的等级主要在等级 3～7，所占比例最大的为等级 5(约 20%)。在空间分布上，等级 1 主要分布在下游的巫溪县南部和巫山县境内，而其他等级分布不具有显著特征。

● 生态环境历史演变

根据表 10.9 可分别计算出 1990 年和 2000 年各评价等级所占比例的对比图和累计比例对比图(图 10－10)。

对比 1990 年和 2000 年大宁河流域生态环境综合评价结果，可以发现，在 1990 年大宁河流域生态环境总体比较好。相比之下，2000 年的评价结果显示生态环境有退化的趋势。如果把 1～3 等级划为Ⅲ类，4～7 等级划为Ⅱ类，8～10 等级划为Ⅰ类，1990 年分别为：Ⅰ类 0.459，Ⅱ类 0.328，Ⅲ类 0.213，而 2000 年分别为：Ⅰ类 0.108，Ⅱ类 0.632，Ⅲ类 0.260。1990 年大宁河流域生态环境总体评价以Ⅰ类为主，将近 50%；而 2000 年总体评价以Ⅱ类为主，超过了 60%。可见，大宁河流域生态环境由 1990 年的Ⅰ类为主，已退化为 2000 年的Ⅱ类为主。相对来说，Ⅲ类生态环境基本上没有多少变化。

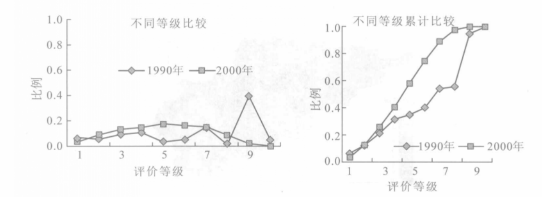
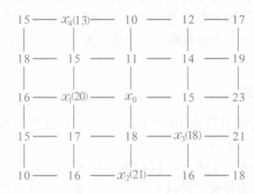


图 10－10　1990 年和 2000 年生态环境综合评价等级对比图

【思考题 10】

1. 什么是区域化变量？什么是协方差函数和变差函数？
2. 什么是克立格方法？举例说明克立格方法在环境科学中的应用。
3. 假设某地区采样分析污染物浓度如下图所示，计算该区域的区域化变量的实验变差函数。



4. 估计第 3 题中球状模型的参数。
5. 试用普通克立格估计第 3 题中 x_0 点的值。
6. 什么是主成分分析？什么是空间主成分分析？
7. 空间主成分分析的步骤有哪些？
8. 空间主成分分析的目的是什么？
9. 在应用 ARC/INFO 进行空间主成分分析时，都涉及到什么操作？
10. 举例说明空间主成分分析的作用。

【参考文献】

[1] 儒尔奈耳 A G，尤日布雷格茨 CH J. 矿业地质统计学 [M]. 侯景儒，黄竞先，杨尔煦，等译. 北京：冶金工业出版社，1982.

[2] 阿格特伯格 F P. 地质数学 [M]. 张中民，译. 北京：科学出版社，1980.

[3] 戴维 M，矿产储量的地质统计学评价(数学地质进展 2)[M]. 孙惠文，刘承祚，译. 北京：地质出版社，1989.

[4] 侯景儒，郭光裕. 矿床统计预测及地质统计学的理论与应用 [M]. 北京：冶金工业出版社，1993.

[5] 侯景儒，黄竞先. 地质统计学及其在矿产储量计算中的应用 [M]. 北京：地质出版社，1982.

[6] 孙洪泉. 地质统计学及其应用 [M]. 徐州：中国矿业大学出版社，1990.

[7] 王仁铎，胡光道. 线性地质统计学 [M]. 北京：地质出版社，1988.

[8] 王政权. 地统计学及在生态学中的应用 [M]. 北京：科学出版社，1999.

[9] 徐建华. 计量地理学 [M]. 北京：高等教育出版社，2006.

[10] 於崇文，蒋耀淞，王长庚，等. 数学地质的方法与应用 [M]. 北京：冶金工业出版社，1980.

[11] CLARK I. Practical geostatistics [M]. London：Applied Science Publication Ltd.，1979.

[12] GUARASCIO M，DAVID M，HUIJBREGTS C. Advanced geostatistics in the mining industry [M]. Dordrecht：Kluwer Academic Publishers，1975.

[13] 黄裕婕，张增祥，周全斌. 西藏中部的生态环境综合评价 [J]. 山地学报，2000，18(4)：31-34.

[14] 王思远，王光谦，陈志祥. 黄河流域生态环境综合评价及其演变 [J]. 山地学报，2004，22(2)：133-139.

[15] 王思远，张增祥，赵晓丽，等. 遥感与 GIS 技术支持下的湖北省生态环境综合分析 [J]. 地球科学进展，2002，17(3)：426-431.

[16] 左伟. 基于 RS，GIS 的区域生态安全综合评价研究 [M]. 北京：测绘出版社，2004.

[17] BELLMANN K. Towards to a system analytical and modeling approach for integration of ecological，hydrological，economical and social components of disturbed regions [J]. Landscape Urban Plan，2000，51 (4)：75.

[18] GESSLER P E，MOORE I D，MCKENZIE N J，et al. Soil-landscape modeling and spatial prediction of soil attributes [J]. International Journal on Geographic Information System，1995，9 (4)：421-432.

[19] WILSON J P，GALLANT J C. Eros-a grid-based program for estimating spatially-distributed erosion indices [J]. Computer Geoscience，1996，22 (7)：707.

部分思考题答案

思考题 1

1. 答案：常用的统计量有样本均值、样本标准差、样本方差、样本的 K 阶原点矩和样本的 K 阶中心矩等。

6. 答案：$\mu=2$，$\sigma=3$

$$P(3<X<9)=\Phi\left(\frac{9-\mu}{\sigma}\right)-\Phi\left(\frac{3-\mu}{\sigma}\right)$$
$$=\Phi\left(\frac{9-2}{3}\right)-\Phi\left(\frac{3-2}{3}\right)=\Phi(2.3333)-\Phi(0.3333)$$
$$=0.990-0.629=0.361$$

7. 答案：$E\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_iE(X_i)=\sum_{i=1}^{n}a_i\mu=\mu$

8. 答案：

$$n=9, \sigma_0=15, \mu_0=500, \alpha=0.01$$
$$\bar{x}=\sum_{i=1}^{9}x_i/n=510.1111$$
$$Z=\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}=2.0222$$
$$z_{0.005}=2.5758$$
$$|Z|<z_{0.005}=2.5758$$

不能拒绝原假设。这台包装机工作正常。

9. 答案：

$$n=7, \mu_0=300, \alpha=0.05$$
$$\bar{x}=\sum_{i=1}^{7}x_i/n=296.2857$$
$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=-0.3596$$
$$t_{0.025}(6)=2.447$$
$$|t|<t_{0.025}(6)=2.447$$

不能拒绝原假设。污灌对玉米穗重无显著影响。

10. 答案：

双因素方差分析表

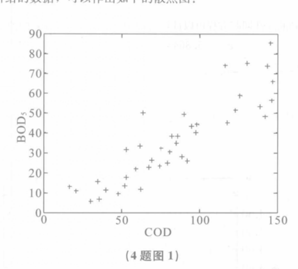
方差来源	SS	自由度	均方差	F 值	
因素 A	0.157 4	2	0.078 7	23.826 7	$F(2,6)=5.14$
因素 B	0.879 6	3	0.293 2	88.756 5	$F(3,6)=4.76$
误差 E	0.019 8	6	0.003 3		
总和 T	1.056 8	11			

大气中飘尘含量的各季节差异显著；大气中飘尘含量的不同区域差异显著。

思考题 2

4. 答案：

(1)根据表中所给的数据，可以作出如下的散点图：



(4 题图 1)

(2)由上面的散点图可以看出，所有的点基本都分布在一条直线周围，故判断 COD 与 BOD_5 之间大致成线性关系。

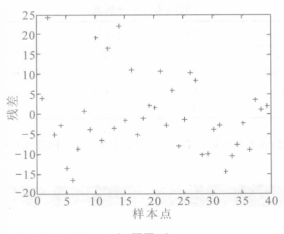
(3)得到的线性回归方程为：$\hat{y}=\hat{a}+\hat{b}x=-5.364\ 2+0.492\ 5x$

(4)因为 $L_{xx}=5.450\ 3\times10^4$，$L_{xy}=2.684\ 3\times10^4$，$L_{yy}=1.685\ 5\times10^4$，

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{2.684\ 3\times10^4}{\sqrt{5.450\ 3\times10^4\times1.685\ 5\times10^4}}=1.492\ 7$$

由此可知 COD 与 BOD_5 的决定系数 $r^2=2.228\ 3$。

(5)根据残差的定义，可以得到以下的残差图：



(4 题图 2)

(6)根据(3)中得到的线性回归方程可以计算得到 COD=99 时，BOD_5 的值为：

$$\hat{y}=\hat{a}+\hat{b}x=-5.364\ 2+0.492\ 5\times 99=43.393\ 0$$

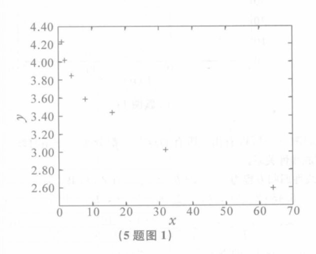
(7)$\hat{\sigma}^2=\dfrac{Q}{n-2}=\dfrac{L_{yy}-b^2L_{xx}}{n-2}=0.280\ 1\times 10^4$

近似有

$$P(43.393\ 0-105.85<y_0<43.393\ 0+105.85)=0.95$$

5. 答案：

(1)散点图：



(5 题图 1)

(2)作变换，对 $y=ax^b$ 两边取对数，得到 $\lg y=\lg a+b\lg x$

令 $y'=\lg y$，$x'=\lg x$，则得 $y'=\lg a+bx'$

对数据进行变换，变换后数据的散点图如下所示：

(5 题图 2)

由上图可知，所有的点基本上都分布在一条直线的周围，故可以采用线性回归分析。

根据变换后的数据，采用一元线性回归分析可以得到：

$$\hat{y}'=\hat{A}+\hat{B}x'=0.6429-0.1107x'$$

将 $y'=\lg y$，$x'=\lg x$ 代入到上式中可以得到：

$$y=4.2160x^{-0.1107}$$

(3)$S_{总}=L_{y'y'}=0.0333$，$S_{回}=0.0311$，$S_{剩}=S_{总}-S_{回}=0.0333-0.0311=0.0022$

则 $F=\dfrac{S_{回}}{S_{剩}/(n-2)}=71.0158$

当 $\alpha=0.05$ 时，查 F 分布表，得到 $F_{\alpha}(1, n-2)=6.61$，由于 $F=71.0158>6.61$，所以认为线性关系式 $y'=A+Bx'=0.6429-0.1107x'$ 显著。

思考题 3

2. 答案：回归方程为：$y=6.0944-0.0371x_1+0.0204x_2-1.1762x_3$。

3. 答案：记铁路客车，铁路里程，公路里程，公路客车分别为 x_1，x_2，x_3，x_4，记旅客周转量为 y。

回归方程为：

$$y=1.5112\times10^4-0.239x_1-0.2533\times10^4x_2-0.0081\times10^4x_3+0.0005\times10^4x_4。$$

4. 答案：

(1) 相关系数矩阵为：

$$\boldsymbol{C}=\begin{pmatrix}1.0000 & 0.3355\\0.3355 & 1.0000\end{pmatrix}$$

(2) 回归方程为：$y=0.6685-0.0294x_1-0.0010x_2$。

(3) 拟合优度检验

$$S_{总}=0.0202,\ S_{残}=0.0061,\ S_{回}=0.0140$$

$$r^2=\frac{S_{回}}{S_{总}}=\frac{0.0140}{0.0202}=0.6955$$

$$r=0.8340$$

在显著性水平 $\alpha=0.05$ 下，$r_\alpha(n-2)=r_\alpha(13)=0.5140$。

由于 $r=0.8340>0.5140=r_\alpha(n-2)$，所以认为回归方程的拟合优度很高。

对每个回归系数的显著性检验：

由程序的运算结果知道：$F_1=581.8442$，$F_2=12.5253$

查 F 分布表得：

$$F_\alpha(1,\ n-k-1)=F_\alpha(1,\ 12)=4.75$$

可以看到：$F_1>F_\alpha$，$F_2>F_\alpha$，说明每个变量对 y 的影响都是很显著的。

思考题 4

1. 答案：(1) 设标准差标准化后的数据矩阵为 $\boldsymbol{X}_1$，$p=4$ 时的明科夫斯基距离矩阵为 $\boldsymbol{A}$。

$$\boldsymbol{X}_1=\begin{pmatrix}
-0.7703 & 0.0374 & -1.2826 & -1.5240 & -0.8805 & -0.3311\\
-0.5974 & -1.2801 & 1.2658 & -0.6232 & -0.6721 & -0.3689\\
-1.2315 & 1.8079 & 0.4902 & 0.1274 & -0.1324 & -0.2771\\
-1.4620 & -0.0861 & -0.2854 & 0.0023 & 1.9968 & -0.3131\\
1.4201 & -0.8272 & -1.3565 & 2.0039 & -0.8221 & -0.2823\\
0.2673 & -0.5802 & 1.2658 & 0.6278 & -0.5822 & 3.1607\\
-0.6550 & 0.1198 & -0.1377 & 1.2533 & -0.4322 & -0.2951\\
-0.0210 & -1.2801 & 1.2658 & -0.6232 & -0.6721 & -0.3689\\
0.4978 & 1.8079 & 0.4902 & 0.1274 & -0.1324 & -0.2771\\
0.8437 & -0.0861 & -0.2854 & 0.0023 & 1.9968 & -0.3131\\
1.7083 & 0.3668 & -1.4303 & -1.3739 & 0.3325 & -0.3337
\end{pmatrix}$$

$$\boldsymbol{A}=\begin{pmatrix}
0.0000 & 2.6022 & 2.2899 & 2.9446 & 3.6551 & 3.8230 & 2.7976 & 2.6066 & 2.3411 & 3.0068 & 2.5137\\
2.6022 & 0.0000 & 3.0958 & 2.7747 & 3.2497 & 3.5480 & 2.1180 & 0.5764 & 3.1065 & 2.8176 & 3.0783\\
2.2899 & 3.0958 & 0.0000 & 2.4108 & 3.3228 & 3.6512 & 1.7783 & 3.1124 & 1.7293 & 2.6895 & 3.1442\\
2.9446 & 2.7747 & 2.4108 & 0.0000 & 3.4985 & 3.7830 & 2.4777 & 2.8176 & 2.6394 & 2.3057 & 3.2679\\
3.6551 & 3.2497 & 3.3228 & 3.4985 & 0.0000 & 3.7281 & 2.1631 & 3.1565 & 2.9249 & 2.9992 & 3.4023\\
3.8230 & 3.5480 & 3.6512 & 3.7830 & 3.7281 & 0.0000 & 3.4858 & 3.5449 & 3.6250 & 3.7416 & 3.8679\\
2.7976 & 2.1180 & 1.7783 & 2.4777 & 2.1631 & 3.4858 & 0.0000 & 2.1222 & 1.8478 & 2.5504 & 3.0090\\
2.6066 & 0.5764 & 3.1124 & 2.8176 & 3.1565 & 3.5449 & 2.1222 & 0.0000 & 3.0950 & 2.7747 & 2.8974\\
2.3411 & 3.1065 & 1.7293 & 2.6394 & 2.9249 & 3.6250 & 1.8478 & 3.0950 & 0.0000 & 2.4110 & 2.2403\\
3.0068 & 2.8176 & 2.6895 & 2.3057 & 2.9992 & 3.7416 & 2.5504 & 2.7747 & 2.4110 & 0.0000 & 1.9196\\
2.5137 & 3.0783 & 3.1442 & 3.2679 & 3.4023 & 3.8679 & 3.0090 & 2.8974 & 2.2403 & 1.9196 & 0.0000
\end{pmatrix}$$

(2) 极差标准化后的数据矩阵为 $\boldsymbol{X}_2$，要素之间的夹角余弦矩阵为 $\boldsymbol{B}$。

$$\boldsymbol{X}_2=\begin{pmatrix} 0.2182 & 0.4267 & 0.0548 & 0.0000 & 0.0000 & 0.0107 \\ 0.2727 & 0.0000 & 1.0000 & 0.2553 & 0.0724 & 0.0000 \\ 0.0727 & 1.0000 & 0.7123 & 0.4681 & 0.2600 & 0.0260 \\ 0.0000 & 0.3867 & 0.4247 & 0.4326 & 1.0000 & 0.0158 \\ 0.9091 & 0.1467 & 0.0274 & 1.0000 & 0.0203 & 0.0246 \\ 0.5455 & 0.2267 & 1.0000 & 0.6099 & 0.1037 & 1.0000 \\ 0.2545 & 0.4533 & 0.4795 & 0.7872 & 0.1558 & 0.0209 \\ 0.4545 & 0.0000 & 1.0000 & 0.2553 & 0.0724 & 0.0000 \\ 0.6182 & 1.0000 & 0.7123 & 0.4681 & 0.2600 & 0.0260 \\ 0.7273 & 0.3867 & 0.4247 & 0.4326 & 1.0000 & 0.0158 \\ 1.0000 & 0.5333 & 0.0000 & 0.0426 & 0.4216 & 0.0100 \end{pmatrix}.$$

$$\boldsymbol{B}=\begin{pmatrix} 1.0000 & 0.2214 & 0.7444 & 0.3171 & 0.4003 & 0.3527 & 0.5337 & 0.2824 & 0.8441 & 0.5029 & 0.7637 \\ 0.2214 & 1.0000 & 0.6064 & 0.4610 & 0.3657 & 0.7419 & 0.6649 & 0.9878 & 0.6456 & 0.5265 & 0.2426 \\ 0.7444 & 0.6064 & 1.0000 & 0.6971 & 0.3871 & 0.5940 & 0.8520 & 0.5829 & 0.9294 & 0.6279 & 0.4533 \\ 0.3171 & 0.4610 & 0.6971 & 1.0000 & 0.3114 & 0.4400 & 0.6644 & 0.4365 & 0.6339 & 0.8611 & 0.4337 \\ 0.4003 & 0.3657 & 0.3871 & 0.3114 & 1.0000 & 0.5308 & 0.7573 & 0.4538 & 0.5991 & 0.6080 & 0.6312 \\ 0.3527 & 0.7419 & 0.5940 & 0.4400 & 0.5308 & 1.0000 & 0.7003 & 0.7555 & 0.6621 & 0.5466 & 0.3730 \\ 0.5337 & 0.6649 & 0.8520 & 0.6644 & 0.7573 & 0.7003 & 1.0000 & 0.6678 & 0.8627 & 0.6931 & 0.4602 \\ 0.2824 & 0.9878 & 0.5829 & 0.4365 & 0.4538 & 0.7555 & 0.6678 & 1.0000 & 0.6787 & 0.5803 & 0.3627 \\ 0.8441 & 0.6456 & 0.9294 & 0.6339 & 0.5991 & 0.6621 & 0.8627 & 0.6787 & 1.0000 & 0.7589 & 0.7177 \\ 0.5029 & 0.5265 & 0.6279 & 0.8611 & 0.6080 & 0.5466 & 0.6931 & 0.5803 & 0.7589 & 1.0000 & 0.7936 \\ 0.7637 & 0.2426 & 0.4533 & 0.4337 & 0.6312 & 0.3730 & 0.4602 & 0.3627 & 0.7177 & 0.7936 & 1.0000 \end{pmatrix}$$

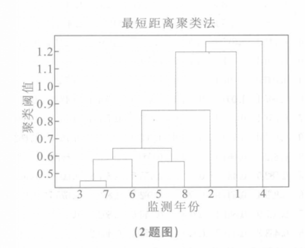
2. 答案：(1)设极差标准化处理后的矩阵为 $\boldsymbol{X}_1$。

$$\boldsymbol{X}_1=\begin{pmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1389 & 0.0000 \\ 0.0000 & 0.6944 & 1.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.2597 & 0.9028 & 0.6154 & 0.3064 & 0.0000 & 0.0000 \\ 0.0649 & 0.9722 & 0.6346 & 0.6774 & 1.0000 & 1.0000 \\ 0.1948 & 0.1389 & 0.2091 & 0.9012 & 0.2778 & 0.0000 \\ 0.0779 & 0.9583 & 0.2933 & 0.1851 & 0.5278 & 0.0000 \\ 0.1039 & 1.0000 & 0.8365 & 0.1887 & 0.3333 & 0.0000 \\ 0.1688 & 0.4583 & 0.4495 & 0.5166 & 0.3889 & 0.0000 \end{pmatrix}$$

(2) 设欧氏距离矩阵为 **A**。

$$A=\begin{pmatrix}0.0000 & 1.8712 & 1.3619 & 2.1030 & 1.2421 & 1.4283 & 1.6051 & 1.1968\\ 1.8712 & 0.0000 & 0.8602 & 1.5228 & 1.0291 & 1.2320 & 0.9488 & 0.8787\\ 1.3619 & 0.8602 & 0.0000 & 1.4768 & 1.0880 & 0.6582 & 0.4556 & 0.6548\\ 2.1030 & 1.5228 & 1.4768 & 0.0000 & 1.5697 & 1.2579 & 1.3139 & 1.3071\\ 1.2421 & 1.0291 & 1.0880 & 1.5697 & 0.0000 & 1.1258 & 1.2861 & 0.5663\\ 1.4283 & 1.2320 & 0.6582 & 1.2579 & 1.1258 & 0.0000 & 0.5791 & 0.6418\\ 1.6051 & 0.9488 & 0.4556 & 1.3139 & 1.2861 & 0.5791 & 0.0000 & 0.7470\\ 1.1968 & 0.8787 & 0.6548 & 1.3071 & 0.5663 & 0.6418 & 0.7470 & 0.0000\end{pmatrix}$$

(3)最短距离聚类图，如下图所示。



(2 题图)

注：监测年份轴中的 1～8 分别对应 1993～2000 年。

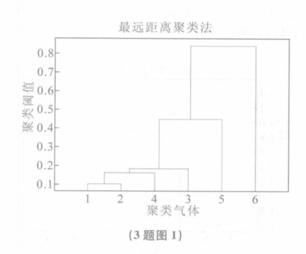
由上图可知，1993，1994 和 1996 年为特殊的年份，不能和其他年份聚类，应各为一类。若其他年份聚类为两类，则 1995，1998 和 1999 年为一类；1997 和 2000 年为一类。

3. 答案：(1)设总和标准化后的数据为 X_1，用夹角余弦求得的 6 种气体间相似系数矩阵为 **A**。

$$X_1=\begin{pmatrix}0.1267 & 0.1239 & 0.0638 & 0.0288 & 0.0178 & 0.0139\\ 0.1109 & 0.0811 & 0.2058 & 0.0835 & 0.0483 & 0.0046\\ 0.0860 & 0.1917 & 0.1626 & 0.1290 & 0.1274 & 0.0272\\ 0.0769 & 0.1401 & 0.1193 & 0.1214 & 0.4395 & 0.0184\\ 0.1900 & 0.0973 & 0.0597 & 0.2428 & 0.0264 & 0.0260\\ 0.1448 & 0.1062 & 0.2058 & 0.1593 & 0.0615 & 0.8738\\ 0.1086 & 0.1313 & 0.1276 & 0.1973 & 0.0835 & 0.0228\\ 0.1561 & 0.1283 & 0.0556 & 0.0379 & 0.1956 & 0.0133\end{pmatrix}$$

$$
A=\begin{pmatrix}
1.0000 & 0.8975 & 0.8211 & 0.8737 & 0.5636 & 0.4418 \\
0.8975 & 1.0000 & 0.8720 & 0.8366 & 0.7365 & 0.3446 \\
0.8211 & 0.8720 & 1.0000 & 0.8166 & 0.6010 & 0.5662 \\
0.8737 & 0.8366 & 0.8166 & 1.0000 & 0.5526 & 0.4433 \\
0.5636 & 0.7365 & 0.6010 & 0.5526 & 1.0000 & 0.1585 \\
0.4418 & 0.3446 & 0.5662 & 0.4433 & 0.1585 & 1.0000
\end{pmatrix}
$$

最远距离聚类图，如下图所示。



(3 题图 1)

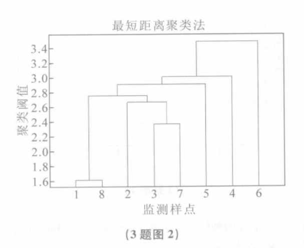
由上图可知，1，2，3 和 4 四种气体，即氯、硫化氢、SO_2 和碳 4 可聚为一类，环氧氯丙烷、环已烷为独立的气体，各为一类。

(2)标准差标准化后的矩阵为 X_2，欧氏距离测度 8 个样点的距离矩阵为 B。

$$
X_2=\begin{pmatrix}
0.0482 & -0.0355 & -1.0473 & -1.3798 & -0.8200 & -0.3923 \\
-0.4020 & -1.4072 & 1.3817 & -0.5960 & -0.5864 & -0.4252 \\
-1.1096 & 2.1404 & 0.6425 & 0.0572 & 0.0187 & -0.3454 \\
-1.3669 & 0.4848 & -0.0968 & -0.0517 & 2.4055 & -0.3767 \\
1.8493 & -0.8869 & -1.1177 & 1.6901 & -0.7545 & -0.3498 \\
0.5628 & -0.6031 & 1.3817 & 0.4926 & -0.4855 & 2.6449 \\
-0.4663 & 0.2010 & 0.0440 & 1.0369 & -0.3175 & -0.3610 \\
0.8845 & 0.1064 & -1.1881 & -1.2492 & 0.5398 & -0.3946
\end{pmatrix}
$$

$$
\boldsymbol{B}=\begin{pmatrix}
0.0000 & 2.9419 & 3.4207 & 3.9172 & 3.6612 & 4.3966 & 2.7579 & 1.6141 \\
2.9419 & 0.0000 & 3.7989 & 3.9933 & 4.1045 & 3.4927 & 2.6689 & 3.4994 \\
3.4207 & 3.7989 & 0.0000 & 3.0105 & 4.9277 & 4.5007 & 2.3678 & 3.6667 \\
3.9172 & 3.9933 & 3.0105 & 0.0000 & 5.1272 & 4.9878 & 3.0840 & 3.3642 \\
3.6612 & 4.1045 & 4.9277 & 5.1272 & 0.0000 & 4.2962 & 2.9177 & 3.4985 \\
4.3966 & 3.4927 & 4.5007 & 4.9878 & 4.2962 & 0.0000 & 3.5855 & 4.5315 \\
2.7579 & 2.6689 & 2.3678 & 3.0840 & 2.9177 & 3.5855 & 0.0000 & 3.0519 \\
1.6141 & 3.4994 & 3.6667 & 3.3642 & 3.4985 & 4.5315 & 3.0519 & 0.0000
\end{pmatrix}
$$

最短距离聚类图，如下图所示。



(3 题图 2)

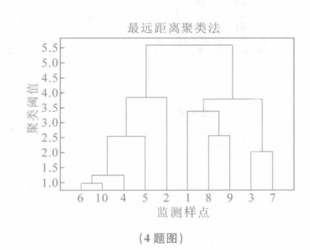
由上图可知，4，5 和 6 号样点比较特殊，不能和其他样点聚类，应各为一类；若其他样点分为两类，则 1 和 8 号样点为一类；2，3 和 7 号为一类。

4. 答案：设标准差标准化后的矩阵为 $\boldsymbol{X}$。

$$
\boldsymbol{X}=\begin{pmatrix}
0.4164 & 1.5098 & 0.4250 & 0.0986 & 2.1733 & 1.1752 & 1.0968 \\
2.6169 & -0.8540 & 0.6559 & -0.7130 & -0.5308 & -0.7456 & -0.6206 \\
-0.3963 & 0.3217 & -0.2539 & 0.2956 & 0.9702 & 0.4892 & 1.4147 \\
-0.6063 & -0.8050 & -0.5975 & -0.6680 & 0.0001 & -0.6084 & -0.7768 \\
-0.8893 & -1.7787 & 0.0596 & -1.0912 & -1.1944 & -0.2954 & -0.8604 \\
-0.4237 & -0.4055 & -0.7826 & -0.5520 & -1.0236 & -0.9557 & -1.1336 \\
-0.3963 & 0.4351 & -1.0274 & -0.0050 & 0.2021 & -0.5098 & 0.0820 \\
-0.0493 & 0.3246 & 1.2455 & 1.1353 & -0.5167 & 2.1014 & 0.8515 \\
0.2429 & 1.2713 & 1.6718 & 2.1710 & 0.3263 & 0.2062 & 0.9853 \\
-0.5150 & -0.0192 & -1.3964 & -0.6713 & -0.4065 & -0.8571 & -1.0388
\end{pmatrix}
$$

欧氏距离矩阵为 $\boldsymbol{A}$。

$$
\boldsymbol{A}=\begin{pmatrix}
0.0000 & 5.0093 & 2.1426 & 4.4104 & 5.6039 & 5.0982 & 3.4191 & 3.3946 & 3.2092 & 4.7394 \\
5.0093 & 0.0000 & 4.4974 & 3.5056 & 3.7878 & 3.4773 & 3.8938 & 4.7470 & 4.8705 & 3.8628 \\
2.1426 & 4.4974 & 0.0000 & 3.0509 & 4.1417 & 3.7529 & 2.0164 & 2.8637 & 3.0355 & 3.4783 \\
4.4104 & 3.5056 & 3.0509 & 0.0000 & 1.7808 & 1.2396 & 1.7304 & 4.3007 & 4.7028 & 1.2489 \\
5.6039 & 3.7878 & 4.1417 & 1.7808 & 0.0000 & 1.9085 & 3.2234 & 4.5418 & 5.4586 & 2.5496 \\
5.0982 & 3.4773 & 3.7529 & 1.2396 & 1.9085 & 0.0000 & 2.0604 & 4.6019 & 4.9353 & 0.9736 \\
3.4191 & 3.8938 & 2.0164 & 1.7304 & 3.2234 & 2.0604 & 0.0000 & 3.8113 & 3.8042 & 1.5962 \\
3.3946 & 4.7470 & 2.8637 & 4.3007 & 4.5418 & 4.6019 & 3.8113 & 0.0000 & 2.5604 & 4.7871 \\
3.2092 & 4.8705 & 3.0355 & 4.7028 & 5.4586 & 4.9353 & 3.8042 & 2.5604 & 0.0000 & 5.0495 \\
4.7394 & 3.8628 & 3.4783 & 1.2489 & 2.5496 & 0.9736 & 1.5962 & 4.7871 & 5.0495 & 0.0000
\end{pmatrix}
$$



（4 题图）

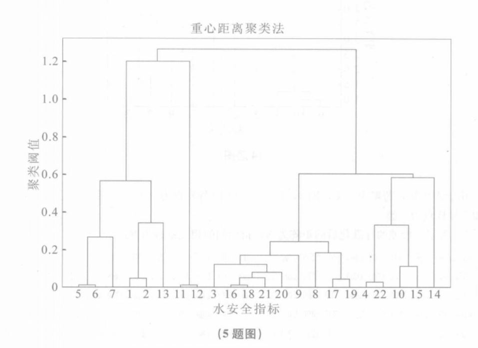
由上图可知，若聚为三类，则 2，4，5，6 和 10 号样点为一类；1，8 和 9 号样点为一类；3 和 7 号样点为一类。

5. 答案：极差的标准化后的矩阵为 $\boldsymbol{X}$，指标间的相关系数为 $\boldsymbol{R}$。

$$
X=\begin{pmatrix}
0.0778 & 0.0000 & 0.0588 & 0.6791 & 0.0397 & 0.1083 & 0.0000 & 0.1509 & 0.3143 & 0.6660 & 1.0000 & 1.0000 & 0.9456 & 0.1017 & 0.8416 & 0.0386 & 0.0000 & 0.0065 & 0.0000 & 0.0000 & 0.0344 & 0.8320 \\
0.5778 & 0.9330 & 0.0933 & 0.0634 & 1.0000 & 1.0000 & 1.0000 & 0.6616 & 0.0000 & 0.6275 & 0.1876 & 0.0000 & 1.0000 & 0.1669 & 0.5674 & 0.0449 & 0.3535 & 0.0000 & 0.2723 & 0.4426 & 0.1868 & 0.0210 \\
0.0272 & 0.0431 & 0.0994 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.7690 & 0.8004 & 1.0000 & 0.0607 & 0.0270 & 0.3026 & 1.0000 & 0.7686 & 0.3360 & 1.0000 & 0.4903 & 1.0000 & 0.5627 & 0.7145 & 1.0000 \\
0.0000 & 0.0526 & 1.0000 & 0.2431 & 0.1190 & 0.2870 & 0.8837 & 1.0000 & 1.0000 & 0.7733 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 1.0000 & 0.7305 & 1.0000 & 0.8195 & 1.0000 & 1.0000 & 0.4763 \\
1.0000 & 1.0000 & 0.0000 & 0.0000 & 0.3810 & 0.4045 & 0.2791 & 0.0000 & 0.2339 & 0.0000 & 0.1634 & 0.0486 & 0.9645 & 0.1907 & 0.0000 & 0.0000 & 0.0183 & 0.0042 & 0.2851 & 0.2039 & 0.0000 & 0.0000
\end{pmatrix}
$$

$$
\boldsymbol{R}=\begin{pmatrix}
1.0000 & 0.9501 & -0.4860 & -0.7306 & 0.6084 & 0.5602 & 0.1525 & -0.6083 & -0.6977 & -0.9015 & -0.1724 & -0.3038 & 0.6685 & -0.2383 & -0.9601 & -0.5916 & -0.5910 & -0.6230 & -0.4497 & -0.4015 & -0.6637 & -0.8458 \\
0.9501 & 1.0000 & -0.4213 & -0.7874 & 0.8183 & 0.7770 & 0.3866 & -0.3960 & -0.7412 & -0.7531 & -0.2767 & -0.4291 & 0.6413 & -0.2453 & -0.8474 & -0.5390 & -0.4689 & -0.5826 & -0.3960 & -0.2599 & -0.5763 & -0.9151 \\
-0.4860 & -0.4213 & 1.0000 & -0.1476 & -0.2447 & -0.0970 & 0.5386 & 0.7017 & 0.7291 & 0.3235 & -0.3942 & -0.2810 & -0.8046 & -0.3603 & 0.5873 & 0.9627 & 0.4543 & 0.9000 & 0.4971 & 0.8504 & 0.8062 & 0.0566 \\
-0.7306 & -0.7874 & -0.1476 & 1.0000 & -0.6989 & -0.7644 & -0.7081 & 0.1778 & 0.4592 & 0.7202 & 0.2729 & 0.3714 & -0.3411 & 0.7224 & 0.5356 & 0.0644 & 0.4822 & 0.1890 & 0.3817 & -0.0664 & 0.2939 & 0.9675 \\
0.6084 & 0.8183 & -0.2447 & -0.6989 & 1.0000 & 0.9864 & 0.6851 & -0.0118 & -0.7296 & -0.3215 & -0.2145 & -0.3750 & 0.5435 & -0.3083 & -0.4024 & -0.3881 & -0.2740 & -0.4634 & -0.3664 & -0.0650 & -0.3904 & -0.8013 \\
0.5602 & 0.7770 & -0.0970 & -0.7644 & 0.9864 & 1.0000 & 0.7836 & 0.0539 & -0.6620 & -0.3136 & -0.2292 & -0.3779 & 0.4663 & -0.4296 & -0.3273 & -0.2651 & -0.2708 & -0.3615 & -0.3570 & 0.0285 & -0.3144 & -0.8280 \\
0.1525 & 0.3866 & 0.5386 & -0.7081 & 0.6851 & 0.7836 & 1.0000 & 0.5261 & -0.0838 & -0.0219 & -0.4809 & -0.5351 & -0.1359 & -0.5312 & 0.1029 & 0.3876 & 0.1121 & 0.2759 & 0.0697 & 0.5854 & 0.2718 & -0.6445 \\
-0.6083 & -0.3960 & 0.7017 & 0.1778 & -0.0118 & 0.0539 & 0.5261 & 1.0000 & 0.6390 & 0.7518 & -0.5879 & -0.5151 & -0.7945 & 0.2070 & 0.6886 & 0.7834 & 0.8680 & 0.8002 & 0.7648 & 0.9074 & 0.8966 & 0.2351 \\
-0.6977 & -0.7412 & 0.7291 & 0.4592 & -0.7296 & -0.6620 & -0.0838 & 0.6390 & 1.0000 & 0.5534 & -0.3859 & -0.2163 & -0.9899 & 0.2860 & 0.6005 & 0.8750 & 0.7484 & 0.9346 & 0.8215 & 0.7285 & 0.9074 & 0.5849 \\
-0.9015 & -0.7531 & 0.3235 & 0.7202 & -0.3215 & -0.3136 & -0.0219 & 0.7518 & 0.5534 & 1.0000 & -0.0353 & 0.0508 & -0.5977 & 0.4626 & 0.8822 & 0.4652 & 0.7587 & 0.5249 & 0.5578 & 0.4571 & 0.6629 & 0.7454 \\
-0.1724 & -0.2767 & -0.3942 & 0.2729 & -0.2145 & -0.2292 & -0.4809 & -0.5879 & -0.3859 & -0.0353 & 1.0000 & 0.9839 & 0.5335 & -0.2845 & 0.1540 & -0.4833 & -0.6447 & -0.5329 & -0.7519 & -0.7621 & -0.5947 & 0.3203 \\
-0.3038 & -0.4291 & -0.2810 & 0.3714 & -0.3750 & -0.3779 & -0.5351 & -0.5151 & -0.2163 & 0.0508 & 0.9839 & 1.0000 & 0.3818 & -0.2487 & 0.2623 & -0.3473 & 0.5488 & -0.3880 & -0.6414 & -0.6740 & -0.4624 & 0.4460 \\
0.6685 & 0.6413 & -0.8046 & -0.3411 & 0.5435 & 0.4663 & -0.1359 & -0.7945 & -0.9899 & -0.5977 & 0.5335 & 0.3818 & 1.0000 & -0.2530 & -0.6230 & -0.9353 & -0.8334 & -0.9826 & -0.8791 & -0.8694 & -0.9809 & -0.4583 \\
-0.2383 & -0.2453 & -0.3603 & 0.7224 & -0.3083 & -0.4296 & -0.5312 & 0.2070 & 0.2860 & 0.4626 & -0.2845 & -0.2487 & -0.2530 & 1.0000 & 0.0344 & -0.0983 & 0.6331 & 0.0750 & 0.6039 & 0.0501 & 0.2562 & 0.5550 \\
-0.9601 & -0.8474 & 0.5873 & 0.5356 & -0.4024 & -0.3273 & 0.1029 & 0.6886 & 0.6005 & 0.8822 & 0.1540 & 0.2623 & -0.6230 & 0.0344 & 1.0000 & 0.6321 & 0.5326 & 0.6215 & 0.3584 & 0.4743 & 0.6555 & 0.6728 \\
-0.5916 & -0.5390 & 0.9627 & 0.0644 & -0.3881 & -0.2651 & 0.3876 & 0.7834 & 0.8750 & 0.4652 & -0.4833 & -0.3473 & -0.9353 & -0.0983 & 0.6321 & 1.0000 & 0.6530 & 0.9843 & 0.6990 & 0.9060 & 0.9286 & 0.2386 \\
-0.5910 & -0.4689 & 0.4543 & 0.4822 & -0.2740 & -0.2708 & 0.1121 & 0.8680 & 0.7484 & 0.7587 & -0.6447 & -0.5488 & -0.8334 & 0.6331 & 0.5326 & 0.6530 & 1.0000 & 0.7538 & 0.9534 & 0.7837 & 0.8869 & 0.4675 \\
-0.6230 & -0.5826 & 0.9000 & 0.1890 & -0.4634 & -0.3615 & 0.2759 & 0.8002 & 0.9346 & 0.5249 & -0.5329 & -0.3880 & -0.9826 & 0.0750 & 0.6215 & 0.9843 & 0.7538 & 1.0000 & 0.8052 & 0.9080 & 0.9697 & 0.3360 \\
-0.4497 & -0.3960 & 0.4971 & 0.3817 & -0.3664 & -0.3570 & 0.0697 & 0.7648 & 0.8215 & 0.5578 & -0.7519 & -0.6414 & -0.8791 & 0.6039 & 0.3584 & 0.6990 & 0.9534 & 0.8052 & 1.0000 & 0.8045 & 0.8916 & 0.3787 \\
-0.4015 & -0.2599 & 0.8504 & -0.0664 & -0.0650 & 0.0285 & 0.5854 & 0.9074 & 0.7285 & 0.4571 & -0.7621 & -0.6740 & -0.8694 & 0.0501 & 0.4743 & 0.9060 & 0.7837 & 0.9080 & 0.8045 & 1.0000 & 0.9273 & 0.0275 \\
-0.6637 & -0.5763 & 0.8062 & 0.2939 & -0.3904 & -0.3144 & 0.2718 & 0.8966 & 0.9074 & 0.6629 & -0.5947 & -0.4624 & -0.9809 & 0.2562 & 0.6555 & 0.9286 & 0.8869 & 0.9697 & 0.8916 & 0.9273 & 1.0000 & 0.3943 \\
-0.8458 & -0.9151 & 0.0566 & 0.9675 & -0.8013 & -0.8280 & -0.6445 & 0.2351 & 0.5849 & 0.7454 & 0.3203 & 0.4460 & -0.4583 & 0.5550 & 0.6728 & 0.2386 & 0.4675 & 0.3360 & 0.3787 & 0.0275 & 0.3943 & 1.0000
\end{pmatrix}
$$

重心距离聚类图，如下图所示。



(5 题图)

思考题 5

1. 答案：设 X 上的模糊集 $\widetilde{A}$= “严重污染程度”，则 $\widetilde{A}$ 可以表示为：

$$\widetilde{A}=0.3/x_1+0.7/x_2+0.9/x_3$$

$$或\ \widetilde{A}=\{(x_1,0.3),(x_2,0.7),(x_3,0.9)\}$$

2. 答案：模糊矩阵 $\boldsymbol{R}=(r_{ij})_{n\times n}$，满足

- 自反性：$r_{ii}=1$
- 对称性：$r_{ij}=r_{ji}$
- 传递性：$\boldsymbol{R}\circ\boldsymbol{R}\subseteq\boldsymbol{R}$

$$\boldsymbol{R}\circ\boldsymbol{R}=\begin{pmatrix}1.0 & 0.5 & 0.9\\ 0.5 & 1.0 & 0.5\\ 0.9 & 0.5 & 1.0\end{pmatrix}\circ\begin{pmatrix}1.0 & 0.5 & 0.9\\ 0.5 & 1.0 & 0.5\\ 0.9 & 0.5 & 1.0\end{pmatrix}=\begin{pmatrix}1.0 & 0.5 & 0.9\\ 0.5 & 1.0 & 0.5\\ 0.9 & 0.5 & 1.0\end{pmatrix}=\boldsymbol{R}$$

则 $\boldsymbol{R}=\begin{pmatrix}1.0 & 0.5 & 0.9\\ 0.5 & 1.0 & 0.5\\ 0.9 & 0.5 & 1.0\end{pmatrix}$ 是一个模糊等价矩阵。

3. 答案：模糊矩阵 $\boldsymbol{R}=(r_{ij})_{n\times n}$，满足

- 自反性：$r_{ii}=1$
- 对称性：$r_{ij}=r_{ji}$
- 传递性：$\boldsymbol{R}\circ\boldsymbol{R}\subseteq\boldsymbol{R}$

经过计算 $\boldsymbol{R}^2\circ\boldsymbol{R}^2=\boldsymbol{R}^2$，故 $\boldsymbol{R}^2$ 就是所求的模糊等价矩阵。

$$\boldsymbol{R}^2=\begin{pmatrix}1.0000 & 0.0000 & 0.9412 & 0.9412\\ 0.0000 & 1.0000 & 0.0000 & 0.0000\\ 0.9412 & 0.0000 & 1.0000 & 0.9412\\ 0.9412 & 0.0000 & 0.9412 & 0.0000\end{pmatrix}$$

4. 答案：

(1)数据标准化

为了使不同量纲的数据也能进行比较，用标准差变换、极差变换等方法对数据进行适当的变换。根据模糊矩阵的要求将数据压缩到区间 [0，1]。

(2)建立模糊相似矩阵

模糊相似矩阵的建立，即标出衡量被分类对象间相似程度的统计量 r_{ij}(i，j=1，2，…，n)，常使用的方法有：相似系数法、距离法、主观评分法等。

(3)聚类分析

将模糊相似矩阵进行聚类分析，常用的方法有：模糊等价矩阵聚类法、直接聚类法等。

5. 答案：

(1)传递闭包法

经计算，$\boldsymbol{R}=\begin{pmatrix} 1.0000 & 0.7942 & 0.8765 & 0.4055 & 0.8856 & 0.4894 & 0.3877 \\ 0.7942 & 1.0000 & 0.6406 & 0.3373 & 0.6617 & 0.5526 & 0.4129 \\ 0.8765 & 0.6406 & 1.0000 & 0.7063 & 0.9112 & 0.6340 & 0.6518 \\ 0.4055 & 0.3373 & 0.7063 & 1.0000 & 0.5582 & 0.8522 & 0.8256 \\ 0.8856 & 0.6617 & 0.9112 & 0.5582 & 1.0000 & 0.5339 & 0.5988 \\ 0.4894 & 0.5526 & 0.6340 & 0.8522 & 0.5339 & 1.0000 & 0.8180 \\ 0.3877 & 0.4129 & 0.6518 & 0.8256 & 0.5988 & 0.8180 & 1.0000 \end{pmatrix}$

$\boldsymbol{R}^4 \circ \boldsymbol{R}^4=\boldsymbol{R}^4$，所以传递闭包为$\boldsymbol{R}=\boldsymbol{R}^4$。

$$\boldsymbol{R}^4=\begin{pmatrix} 1.0000 & 0.7942 & 0.8856 & 0.7063 & 0.8856 & 0.7063 & 0.7063 \\ 0.7942 & 1.0000 & 0.7942 & 0.7063 & 0.7942 & 0.7063 & 0.7063 \\ 0.8856 & 0.7942 & 1.0000 & 0.7063 & 0.9112 & 0.7063 & 0.7063 \\ 0.7063 & 0.7063 & 0.7063 & 1.0000 & 0.7063 & 0.8522 & 0.8256 \\ 0.8856 & 0.7942 & 0.9112 & 0.7063 & 1.0000 & 0.7063 & 0.7063 \\ 0.7063 & 0.7063 & 0.7063 & 0.8522 & 0.7063 & 1.0000 & 0.8256 \\ 0.7063 & 0.7063 & 0.7063 & 0.8256 & 0.7063 & 0.8256 & 1.0000 \end{pmatrix}$$

● 当$\lambda \geqslant 1.0000$时

此时只有对角线元素大于等于1，故对角线元素全变成1，其余全部为0，成为单位矩阵，共分为7类：$\{u_1\}$，$\{u_2\}$，$\{u_3\}$，$\{u_4\}$，$\{u_5\}$，$\{u_6\}$，$\{u_7\}$ 使把每一个元素分为一类，是最细的分类。

● 当$\lambda \geqslant 0.8856$时

此时小于0.8856的元素都变成0，大于等于0.8856的元素变成1，即有：

$$\boldsymbol{R}_\lambda=\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

可以看出共分5类：$\{u_1, u_3, u_5\}$，$\{u_2\}$，$\{u_4\}$，$\{u_6\}$，$\{u_7\}$。

● 当$\lambda \geqslant 0.7942$时

此时小于0.7942的元素都变成0，大于等于0.7942的元素变成1，即有：

$$\boldsymbol{R}_\lambda=\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

可以看出共分 2 类：$\{u_1, u_2, u_3, u_5\}$，$\{u_4, u_6, u_7\}$。

● 当 $\lambda \geqslant 0.0634$ 时

矩阵的所有元素都变成 1，只分成 1 类，是最粗的分类。

(2)直接聚类法

$$R=\begin{pmatrix} 1.0 & 0.8 & 1.0 & 0.2 & 0.8 & 0.5 & 0.3 \\ 0.8 & 1.0 & 0.4 & 0.3 & 0.7 & 0.6 & 0.3 \\ 1.0 & 0.4 & 1.0 & 0.7 & 1.0 & 0.6 & 0.5 \\ 0.2 & 0.3 & 0.7 & 1.0 & 0.5 & 0.8 & 0.6 \\ 0.8 & 0.7 & 1.0 & 0.5 & 1.0 & 0.2 & 0.7 \\ 0.5 & 0.6 & 0.6 & 0.8 & 0.2 & 1.0 & 0.8 \\ 0.3 & 0.3 & 0.5 & 0.6 & 0.7 & 0.8 & 1.0 \end{pmatrix}$$

$\boldsymbol{R}$ 为论域上的模糊相似矩阵。

● 取 $\boldsymbol{R}$ 中的最大值 $\lambda_1=1$，$r_{13}=r_{35}$，这样，在 $\lambda_1=1$ 水平上的等价类为：$\{u_1, u_3, u_5\}$，$\{u_2\}$，$\{u_4\}$，$\{u_6\}$，$\{u_7\}$。

● 取 $\boldsymbol{R}$ 中的次大值 $\lambda_2=0.8$，由于 $r_{12}=r_{15}=r_{46}=r_{67}=0.8$，故相似类为：$\{u_1, u_2, u_3, u_5\}$，$\{u_4, u_6, u_7\}$。

● 取 $\boldsymbol{R}$ 中的第三大值 $\lambda_3=0.7$，由于 $r_{34}=r_{25}=r_{57}=0.7$，所有元素只分为一类。

6. 答案：利用传递闭包法：

$$R=\begin{pmatrix} 1.0000 & 0.9321 & 0.6744 & 0.9321 & 0.6744 & 0.6744 & 0.6744 & 0.9321 & 0.9321 \\ 0.9321 & 1.0000 & 0.6744 & 0.9359 & 0.6744 & 0.6744 & 0.6744 & 0.9550 & 0.9359 \\ 0.6744 & 0.6744 & 1.0000 & 0.6744 & 0.7959 & 0.7959 & 0.7959 & 0.6744 & 0.6744 \\ 0.9321 & 0.9359 & 0.6744 & 1.0000 & 0.6744 & 0.6744 & 0.6744 & 0.9359 & 0.9471 \\ 0.6744 & 0.6744 & 0.7959 & 0.6744 & 1.0000 & 0.9903 & 0.9895 & 0.6744 & 0.6744 \\ 0.6744 & 0.6744 & 0.7959 & 0.6744 & 0.9903 & 1.0000 & 0.9895 & 0.6744 & 0.6744 \\ 0.6744 & 0.6744 & 0.7959 & 0.6744 & 0.9895 & 0.9895 & 1.0000 & 0.6744 & 0.6744 \\ 0.9321 & 0.9550 & 0.6744 & 0.9359 & 0.6744 & 0.6744 & 0.6744 & 1.0000 & 0.9359 \\ 0.9321 & 0.9359 & 0.6744 & 0.9471 & 0.6744 & 0.6744 & 0.6744 & 0.9359 & 1.0000 \end{pmatrix}$$

● 当 $\lambda \geqslant 1.0000$ 时

此时只有对角线元素大于等于 1，故对角线元素全变成 1，其余全部为 0，成为单位矩阵，共分为 9 类：$\{u_1\}$，$\{u_2\}$，$\{u_3\}$，$\{u_4\}$，$\{u_5\}$，$\{u_6\}$，$\{u_7\}$，$\{u_8\}$，$\{u_9\}$ 使把每一个元素分为一类，是最细的分类。

● 当 $\lambda \geqslant 0.9321$ 时

$$
R_{\lambda}=\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}
$$

可以看出共分 3 类：$\{u_1, u_2, u_4, u_8, u_9\}$，$\{u_3\}$，$\{u_5, u_6, u_7\}$。

● 当 $\lambda \geqslant 0.7942$ 时

$$
R_{\lambda}=\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1\\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}
$$

可以看出共分 2 类：$\{u_1, u_2, u_4, u_8, u_9\}$，$\{u_3, u_5, u_6, u_7\}$。

● 当 $\lambda \geqslant 0.6744$ 时

矩阵的所有元素都变成 1，只分成 1 类，是最粗的分类。

7. 答案：利用传递闭包法：

$$
R=\begin{pmatrix}
0.8697 & 0.9601 & 0.9503 & 0.7241 & 0.7368 & 0.8094 & 0.8486 & 0.9141 & 0.9670\\
0.6622 & 0.7298 & 0.8361 & 0.8785 & 0.8522 & 0.4865 & 0.5762 & 0.6627 & 0.8287\\
0.6749 & 0.8290 & 0.8536 & 0.6906 & 0.8592 & 0.6460 & 0.6438 & 0.8461 & 0.8337\\
0.9463 & 0.8074 & 0.9059 & 0.5886 & 0.6266 & 0.7724 & 0.8469 & 0.5571 & 0.6898\\
0.6547 & 0.8782 & 0.7380 & 0.5620 & 0.6633 & 0.7985 & 0.7287 & 0.9921 & 0.8315\\
0.6428 & 0.7042 & 0.7071 & 0.6108 & 0.3891 & 0.4860 & 0.6089 & 0.6696 & 0.8821\\
0.2301 & 0.4542 & 0.3548 & 0.8049 & 0.7775 & 0.3462 & 0.2692 & 0.6380 & 0.4938\\
0.7398 & 0.7852 & 0.8931 & 0.7374 & 0.7631 & 0.5489 & 0.6432 & 0.6969 & 0.8729\\
0.8839 & 0.9797 & 0.9593 & 0.7061 & 0.6891 & 0.8408 & 0.8885 & 0.9181 & 0.9713\\
0.8888 & 0.9739 & 0.9728 & 0.7215 & 0.7176 & 0.8250 & 0.8783 & 0.9013 & 0.9690\\
0.5994 & 0.7853 & 0.6770 & 0.5296 & 0.8124 & 0.7438 & 0.6128 & 0.9151 & 0.6763\\
1.0000 & 0.9201 & 0.9397 & 0.5860 & 0.5836 & 0.9138 & 0.9670 & 0.7079 & 0.7763\\
0.9201 & 1.0000 & 0.9605 & 0.6543 & 0.6798 & 0.9255 & 0.9467 & 0.9159 & 0.9105\\
0.9397 & 0.9605 & 1.0000 & 0.6962 & 0.6923 & 0.8417 & 0.9169 & 0.8069 & 0.9096\\
0.5860 & 0.6543 & 0.6962 & 1.0000 & 0.8527 & 0.4796 & 0.5402 & 0.6167 & 0.7219\\
0.5836 & 0.6798 & 0.6923 & 0.8527 & 1.0000 & 0.5706 & 0.5172 & 0.7163 & 0.6393\\
0.9138 & 0.9255 & 0.8417 & 0.4796 & 0.5706 & 1.0000 & 0.9608 & 0.8121 & 0.6956\\
0.9670 & 0.9467 & 0.9169 & 0.5402 & 0.5172 & 0.9608 & 1.0000 & 0.7630 & 0.7797\\
0.7079 & 0.9159 & 0.8069 & 0.6167 & 0.7163 & 0.8121 & 0.7630 & 1.0000 & 0.8790\\
0.7763 & 0.9105 & 0.9096 & 0.7219 & 0.6393 & 0.6956 & 0.7797 & 0.8790 & 1.0000
\end{pmatrix}
$$

- 当$\lambda \geqslant 1.00$时

此时只有对角线元素大于等于1，故对角线元素全变成1，其余全部为0，成为单位矩阵，共分为20类，每一个元素分为一类，是最细的分类。

- 当$\lambda \geqslant 0.95$时

可以看出共分11类：$\{u_1, u_9, u_{10}, u_{13}, u_{14}, u_{20}\}$，$\{u_2, u_8\}$，$\{u_3\}$，$\{u_4\}$，$\{u_5, u_{19}\}$，$\{u_6\}$，$\{u_7\}$，$\{u_{11}\}$，$\{u_{12}, u_{17}, u_{18}\}$，$\{u_{15}\}$，$\{u_{16}\}$。

- 当$\lambda \geqslant 0.90$时

可以看出共分6类：$\{u_1, u_4, u_5, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{17}, u_{18}, u_{19}, u_{20}\}$，$\{u_2, u_3, u_8\}$，$\{u_6\}$，$\{u_7\}$，$\{u_{15}\}$，$\{u_{16}\}$。

- 当$\lambda \geqslant 0.85$时

可以看出共分2类：$\{u_1, u_2, u_3, u_4, u_5, u_6, u_8, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}, u_{17}, u_{18}, u_{19}, u_{20}\}$，$\{u_7\}$。

- 当$\lambda \geqslant 0.80$时

矩阵的所有元素都变成1，只分成1类，是最粗的分类。

思考题 6

4. 答案：

(1)距离判别方法

①将一级标准记为G_1，二级标准记为G_2。经过计算，各类样本的指标均值为：

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{x}_i = \bar{\boldsymbol{x}} = (0.0250 \quad 0.0275 \quad 0.0639)'$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{y}_i = \bar{\boldsymbol{y}} = (0.1149 \quad 0.0295 \quad 0.2190)'$$

$$\bar{\boldsymbol{\mu}} = (0.0699 \quad 0.0285 \quad 0.1415)'$$

总体协方差的逆矩阵为：

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.0005 & -0.0001 & -0.0006 \\ -0.0001 & 0.0007 & 0.0003 \\ -0.0006 & 0.0003 & 0.0024 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1} = 10^3 \times \begin{bmatrix} 2.4708 & 0.1506 & 0.5578 \\ 0.1506 & 1.5155 & -0.1308 \\ 0.5578 & -0.1308 & 0.5577 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) = (-308.7812, 3.6016, -136.3149)',$$

从而判别函数：

$$W(x)=(x-\bar{\mu})'\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)$$
$$=-308.7812(x_1-0.0699)+3.6016(x_2-0.0285)-136.3149(x_3-0.1415)$$

将5个待判的样本数据分别代入到上面的判别函数中，可以分别求得函数值为：

$$W_1=-5.5704,\ W_2=12.1639,\ W_3=3.1021,$$
$$W_4=17.8225,\ W_5=32.6918$$

$W_1<0$，$W_2>0$，$W_3>0$，$W_4>0$，$W_5>0$ 根据判别函数的定义，可以判定样本1属于 G_2，样本2～5属于 G_1。

②将二级标准记为 G_1，三级标准记为 G_2。经过计算，各类样本的指标均值为：

$$\hat{\mu}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x_i=\bar{x}=(0.1149\quad 0.0295\quad 0.2190)'$$

$$\hat{\mu}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}y_i=\bar{y}=(0.2212\quad 0.1261\quad 0.3854)'$$

$$\bar{\mu}=(0.1681\quad 0.0778\quad 0.3022)'$$

总体协方差的逆矩阵为：

$$\hat{\Sigma}=\begin{pmatrix}0.0008 & 0.0002 & -0.0004\\ 0.0002 & 0.0003 & 0.0001\\ -0.0004 & 0.0001 & 0.0028\end{pmatrix}$$

$$\hat{\Sigma}^{-1}=10^3\times\begin{pmatrix}1.5901 & -1.0052 & 0.2360\\ -1.0052 & 3.9423 & -0.2432\\ 0.2360 & -0.2432 & 0.3986\end{pmatrix}$$

$$\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)=(-111.3998,\ -233.1156,\ -67.9733)',$$

从而判别函数：

$$W(x)=(x-\bar{\mu})'\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)$$
$$=-111.3998(x_1-0.1681)-233.1156(x_2-0.0778)-67.9733(x_3-0.3022)$$

将5个待判的样本数据分别代入到上面的判别函数中，可以分别求得函数值为：

$$W_1=29.3751,\ W_2=39.2857,\ W_3=35.5357,$$
$$W_4=40.2501,\ W_5=49.2665$$

$W_1>0$，$W_2>0$，$W_3>0$，$W_4>0$，$W_5>0$ 根据判别函数的定义，可以判定样本1～5属于 G_1，即样本1～5不属于三级。

综合①，②判定样本1属于 G_2，样本2～5属于 G_1。

(2)Fisher 判别方法

①将一级标准记为 G_1，二级标准记为 G_2。

根据距离判别中的分析数据，可以得到：

$$y_0=\frac{1}{2}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)=-40.7780$$

$$y=\boldsymbol{c}'\boldsymbol{x}=(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}=-308.7812x_1+3.6016x_2-136.3149x_3$$

$$\overline{y^{(1)}}=\frac{1}{n_1}\sum_{j=1}^{n_1}y_j^{(1)}=-16.3428$$

$$\overline{y^{(2)}}=\frac{1}{n_2}\sum_{j=1}^{n_2}y_j^{(2)}=-65.2132$$

$$\text{即}\overline{y^{(1)}}>y_0>\overline{y^{(2)}}$$

将样本 1，2，3，4，5 的数据分别代入到判别函数中，得到：

$$y_1=-46.3484,\ y_2=-28.6140,\ y_3=-37.6758,$$

$$y_4=-22.9555,\ y_5=-8.0862$$

根据 Fisher 判别准则，可以判定样本 1 属于 G_2，样本 2～5 属于 G_1。这个结果和距离判别的结果是一致的。

②将二级标准记为 G_1，三级标准记为 G_2。

根据距离判别中的分析数据，可以得到：

$$y_0=\frac{1}{2}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)=-57.3993$$

$$y=\boldsymbol{c}'\boldsymbol{x}=(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}=-111.3998x_1-233.1156x_2-67.9733x_3$$

$$\overline{y^{(1)}}=\frac{1}{n_1}\sum_{j=1}^{n_1}y_j^{(1)}=-34.5678$$

$$\overline{y^{(2)}}=\frac{1}{n_2}\sum_{j=1}^{n_2}y_j^{(2)}=-80.2309$$

$$\text{即}\overline{y^{(1)}}>y_0>\overline{y^{(2)}}$$

将样本 1，2，3，4，5 的数据分别代入到判别函数中，得到：

$$y_1=-28.0242,\ y_2=-18.1136,\ y_3=-21.8637,$$

$$y_4=-17.1492,\ y_5=-8.1329$$

根据 Fisher 判别准则，可以判定样本 1～5 属于 G_1，即二级。

综合①，②判定样本 1 属于 G_2，样本 2～5 属于 G_1。

这个结果和距离判别的结果是一致的。

5. 答案：

根据表中的数据，得到：

$$\hat{\boldsymbol{\mu}}_1=(4.2800\quad 22.1400\quad 12.0233\quad 51.1833)'$$

$$\hat{\boldsymbol{\mu}}_2=(1.0700\quad 5.2725\quad 2.7300\quad 20.6750)'$$

$$\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2=(3.2100\quad 16.8675\quad 9.2933\quad 30.5083)'$$

$$\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2=(5.3500\quad 27.4125\quad 14.7533\quad 71.8583)'$$

$$\bar{\boldsymbol{\mu}}=(2.6750\quad 13.7062\quad 7.3767\quad 35.9292)'$$

$$\hat{\boldsymbol{\Sigma}}=\begin{pmatrix} 0.7725 & -0.1690 & 0.5781 & 3.1787 \\ -0.1690 & 28.6615 & 13.3399 & -4.9706 \\ 0.5781 & 13.3399 & 8.4187 & 20.8874 \\ 3.1787 & -4.9706 & 20.8874 & 272.4658 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}=\begin{pmatrix} 13.1850 & 10.8171 & -22.4175 & 1.7621 \\ 10.8171 & 9.9208 & -20.4975 & 1.6261 \\ -22.4175 & -20.4975 & 42.5002 & -3.3705 \\ 1.7621 & 1.6261 & -3.3705 & 0.2712 \end{pmatrix}$$

$$y_0=\frac{1}{2}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)=460.6988$$

$$\begin{aligned} y&=\boldsymbol{c}'\boldsymbol{x}=(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} \\ &=70.2051x_1+61.1830x_2-125.5620x_3+10.0348x_4 \end{aligned}$$

$$\overline{y^{(1)}}=\frac{1}{n_1}\sum_{j=1}^{n_1}y_j^{(1)}=659.0068$$

$$\overline{y^{(2)}}=\frac{1}{n_2}\sum_{j=1}^{n_2}y_j^{(2)}=262.3909$$

即$\overline{y^{(1)}}>y_0>\overline{y^{(2)}}$

将样本 A，B 的数据分别代入到判别函数中，得到：

$$y_1=561.5964,\ y_2=445.0748$$

根据 Fisher 判别准则，样本 A 属于甲地，样本 B 属于 B 地。

思考题 7

3. 答案：

原决策矩阵：

$$\boldsymbol{X}=\begin{pmatrix} 2.0000 & 0.0100 & 0.2000 \\ 4.0000 & 0.0250 & 0.5000 \\ 6.0000 & 0.0500 & 1.0000 \\ 10.0000 & 0.1000 & 1.5000 \\ 15.0000 & 0.2000 & 2.0000 \\ 7.1000 & 0.1440 & 7.0000 \\ 5.7000 & 0.1020 & 5.2700 \\ 5.4000 & 0.1070 & 3.3300 \\ 4.0000 & 0.0560 & 1.7100 \\ 4.2000 & 0.0590 & 1.9000 \\ 4.7000 & 0.0780 & 2.8200 \end{pmatrix}$$

标准化后处理的矩阵为：

$$A=\begin{pmatrix} -1.1751 & -1.3691 & -1.1040 \\ -0.6143 & -1.0939 & -0.9584 \\ -0.0535 & -0.6354 & -0.7159 \\ 1.0681 & 0.2818 & -0.4733 \\ 2.4701 & 2.1162 & -0.2307 \\ 0.2549 & 1.0889 & 2.1952 \\ -0.1376 & 0.3185 & 1.3558 \\ -0.2218 & 0.4102 & 0.4146 \\ -0.6143 & -0.5253 & -0.3714 \\ -0.5582 & -0.4703 & -0.2792 \\ -0.4180 & -0.1217 & 0.1672 \end{pmatrix}$$

计算标准化数据矩阵 $\boldsymbol{A}$ 的相关矩阵。

$$R=\begin{pmatrix} 1.0000 & 0.8622 & 0.1444 \\ 0.8622 & 1.0000 & 0.5692 \\ 0.1444 & 0.5692 & 1.0000 \end{pmatrix}$$

求 $\boldsymbol{R}$ 的特征值 $\lambda_1\geqslant\lambda_2\geqslant\lambda_3$，以及对应的特征向量 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$，$\boldsymbol{u}_3$，要求它们是标准正交的。

$$U=\begin{pmatrix} 0.3466 & -0.3783 & -0.3587 \\ 0.4011 & -0.0394 & 0.4379 \\ 0.2523 & 0.5822 & -0.2034 \end{pmatrix}=(\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \boldsymbol{u}_3)$$

$$\lambda_1=2.1030 \quad \lambda_2=0.8676 \quad \lambda_3=0.0294$$

累计贡献率

第一、第二和第三个成分的累计贡献率分别为：

0.7010　　0.9902　　1.0000

求第一主成分 F_1，有：

$F_1=\boldsymbol{A}\boldsymbol{u}_1=(-1.2350\quad -0.8935\quad -0.4540\quad 0.3638\quad 1.6466\quad 1.0791\quad 0.4222\quad 0.1923\quad -0.5173\quad -0.4525\quad -0.1515)$

综合评价

$F_1=(-1.2350\quad -0.8935\quad -0.4540\quad 0.3638\quad 1.6466\quad 1.0791\quad 0.4222\quad 0.1923\quad -0.5173\quad -0.4525\quad -0.1515)$

按 F_1 由小到大的顺序排列方案的优先次序，结果是：

$$e_1<e_2<e_9<e_3<e_{10}<e_{11}<e_8<e_4<e_7<e_6<e_5$$

根据上面的计算结果：各地区环境质量从优到劣依次为：东部沿岸区、湖心区、全部平均、西部沿岸区、梅梁湖、五里湖。因为第一主成分是高锰酸盐指数，所以根据地表水环境质量标准，可以判断：东部沿岸区为Ⅱ类，湖心区、全部平均、西部沿岸区、梅梁湖为Ⅲ类，

五里湖为Ⅳ类。

4. 答案：原决策矩阵：

$$
\boldsymbol{X}=\begin{pmatrix}
0.221 & 89.52 & 42.66 & 32.63 & 46.50 & 0.530 & 10.90 \\
0.462 & 57.21 & 46.49 & 25.42 & 27.35 & 0.082 & 7.82 \\
0.132 & 73.28 & 31.40 & 34.38 & 37.98 & 0.370 & 11.47 \\
0.109 & 57.88 & 25.70 & 25.82 & 31.11 & 0.114 & 7.54 \\
0.078 & 44.57 & 36.60 & 22.06 & 22.65 & 0.187 & 7.39 \\
0.129 & 63.34 & 22.63 & 26.85 & 23.86 & 0.033 & 6.90 \\
0.132 & 74.83 & 18.57 & 31.71 & 32.54 & 0.137 & 9.08 \\
0.170 & 73.32 & 56.27 & 41.84 & 27.45 & 0.746 & 10.46 \\
0.202 & 86.26 & 63.34 & 51.04 & 33.42 & 0.304 & 10.70 \\
0.119 & 68.62 & 12.45 & 25.79 & 28.23 & 0.056 & 7.07 \\
0.063 & 35.39 & 13.58 & 16.17 & 18.29 & 0.167 & 12.15 \\
0.142 & 68.41 & 29.18 & 33.33 & 28.96 & 0.072 & 10.67 \\
0.134 & 85.39 & 26.60 & 37.90 & 40.04 & 0.290 & 6.60 \\
0.051 & 24.61 & 10.69 & 13.80 & 17.65 & 0.184 & 8.49 \\
0.038 & 42.23 & 5.51 & 10.20 & 11.24 & 0.036 & 5.08 \\
0.121 & 49.73 & 37.14 & 32.78 & 21.41 & 0.579 & 5.62 \\
0.047 & 26.93 & 8.79 & 10.64 & 15.71 & 0.029 & 10.49 \\
0.065 & 60.17 & 13.86 & 18.48 & 21.04 & 0.091 & 10.07 \\
0.065 & 35.84 & 11.64 & 17.23 & 21.37 & 0.055 & 8.50 \\
0.044 & 34.19 & 15.69 & 12.97 & 9.80 & 0.031 & 9.86 \\
0.055 & 30.19 & 9.96 & 13.42 & 13.03 & 0.046 & 10.09 \\
0.058 & 27.78 & 10.99 & 15.65 & 14.19 & 0.034 & 13.08
\end{pmatrix}
$$

标准化后处理的矩阵为：

$$
\boldsymbol{A}=\begin{pmatrix}
1.0946 & 1.6515 & 1.0862 & 0.6905 & 2.2475 & 1.6848 & 0.8432 \\
3.7029 & 0.1064 & 1.3216 & 0.0376 & 0.2715 & -0.5331 & -0.5935 \\
0.1313 & 0.8749 & 0.3941 & 0.8490 & 1.3683 & 0.8927 & 1.1091 \\
-0.1176 & 0.1384 & 0.0438 & 0.0738 & 0.6594 & -0.3747 & -0.7241 \\
-0.4531 & -0.4981 & 0.7137 & -0.2667 & -0.2135 & -0.0133 & -0.7940 \\
0.0989 & 0.3995 & -0.1449 & 0.1671 & -0.0886 & -0.7757 & -1.0226 \\
0.1313 & 0.9490 & -0.3945 & 0.6072 & 0.8070 & -0.2608 & -0.0057 \\
0.5426 & 0.8768 & 1.9227 & 1.5246 & 0.2818 & 2.7542 & 0.6380 \\
0.8889 & 1.4956 & 2.3572 & 2.3578 & 0.8978 & 0.5660 & 0.7499 \\
-0.0093 & 0.6520 & -0.7706 & 0.0711 & 0.3623 & -0.6618 & -0.9433 \\
-0.6154 & -0.9371 & -0.7012 & -0.8001 & -0.6634 & -0.1123 & 1.4263 \\
0.2396 & 0.6420 & 0.2576 & 0.7539 & 0.4376 & -0.5826 & 0.7359 \\
0.1530 & 1.4540 & 0.0991 & 1.1678 & 1.5809 & 0.4966 & -1.1625 \\
-0.7453 & -1.4526 & -0.8788 & -1.0147 & -0.7294 & -0.0281 & -0.2809 \\
-0.8860 & -0.6100 & -1.1972 & -1.3408 & -1.3908 & -0.7608 & -1.8716 \\
0.0123 & -0.2513 & 0.7469 & 0.7041 & -0.3414 & 1.9274 & -1.6197 \\
-0.7886 & -1.3417 & -0.9956 & -1.3009 & -0.9296 & -0.7955 & 0.6520 \\
-0.5938 & 0.2479 & -0.6840 & -0.5909 & -0.3796 & -0.4885 & 0.4561 \\
-0.5938 & -0.9156 & -0.8204 & -0.7041 & -0.3456 & -0.6668 & -0.2763 \\
-0.8211 & -0.9945 & -0.5715 & -1.0899 & -1.5394 & -0.7856 & 0.3581 \\
-0.7020 & -1.1858 & -0.9237 & -1.0492 & -1.2061 & -0.7113 & 0.4654 \\
-0.6695 & -1.3010 & -0.8604 & -0.8472 & -1.0864 & -0.7707 & 1.8601
\end{pmatrix}
$$

计算标准化数据矩阵 **A** 的协方差矩阵。

$$
C=\begin{pmatrix}
1.0000 & 0.5256 & 0.6900 & 0.5207 & 0.5351 & 0.2680 & -0.0611\\
0.5256 & 1.0000 & 0.6520 & 0.8716 & 0.8920 & 0.4949 & -0.0565\\
0.6900 & 0.6520 & 1.0000 & 0.8385 & 0.5939 & 0.6953 & 0.0583\\
0.5207 & 0.8716 & 0.8385 & 1.0000 & 0.7852 & 0.6553 & 0.0234\\
0.5351 & 0.8920 & 0.5939 & 0.7852 & 1.0000 & 0.5051 & 0.0297\\
0.2680 & 0.4949 & 0.6953 & 0.6553 & 0.5051 & 1.0000 & 0.0548\\
-0.0611 & -0.0565 & 0.0583 & 0.0234 & 0.0297 & 0.0548 & 1.0000
\end{pmatrix}
$$

求 $\boldsymbol{C}$ 的特征值 $\lambda_1\geqslant\lambda_2\geqslant\lambda_3\geqslant\lambda_4\geqslant\lambda_5\geqslant\lambda_6\geqslant\lambda_7$，以及对应的特征向量 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$，$\boldsymbol{u}_3$，$\boldsymbol{u}_4$，$\boldsymbol{u}_5$，$\boldsymbol{u}_6$，$\boldsymbol{u}_7$，要求它们是标准正交的。

$$
\boldsymbol{U}=\begin{pmatrix}
0.1379 & -0.1224 & 0.3607 & 0.1801 & 0.1351 & -0.1114 & 0.1125\\
0.1788 & -0.0673 & 0.0185 & -0.1950 & -0.0532 & -0.2802 & -0.2265\\
0.1772 & 0.0467 & 0.0030 & 0.1950 & -0.1694 & 0.1865 & -0.2666\\
0.1878 & 0.0202 & -0.0664 & -0.0459 & -0.2304 & 0.0021 & 0.3518\\
0.1731 & -0.0161 & 0.0486 & -0.2082 & 0.1997 & 0.2737 & 0.0065\\
0.1424 & 0.1259 & -0.3506 & 0.1418 & 0.2079 & -0.1071 & 0.0354\\
0.0028 & 0.6014 & 0.1522 & -0.0340 & 0.0042 & -0.0389 & -0.0008
\end{pmatrix}
$$

$=(\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \boldsymbol{u}_3,\ \boldsymbol{u}_4,\ \boldsymbol{u}_5,\ \boldsymbol{u}_6,\ \boldsymbol{u}_7)$

$$\lambda_1=4.2240\quad \lambda_2=1.0348\quad \lambda_3=0.7226\quad \lambda_4=0.6498$$
$$\lambda_5=0.2407\quad \lambda_6=0.0782\quad \lambda_7=0.0500$$

累计贡献率

各个成分的累计贡献率分别为：

0.603 4　0.751 3　0.854 5　0.947 3　0.981 7　0.992 9　1.000 0

求第一主成分 F_1，有

$F_1=\boldsymbol{A}\boldsymbol{u}_1=$

(1.399 7　0.740 2　0.770 9　0.088 8　−0.116 3　−0.038 0　0.334 4　1.301 4　1.488 5　−0.042 2　−0.653 7　0.329 9　0.859 0　−0.839 8　−1.049 5　0.432 2　−1.041 7　−0.403 7　−0.678 7　−0.974 3　−0.978 2　−0.929 0)

综合评价

$F_1=$(1.399 7　0.740 2　0.770 9　0.088 8　−0.116 3　−0.038 0　0.334 4　1.301 4　1.488 5　−0.042 2　−0.653 7　0.329 9　0.859 0　−0.839 8　−1.049 5　0.432 2　−1.041 7　−0.403 7　−0.678 7　−0.974 3　−0.978 2　−0.929 0)

按 F_1 由大到小的顺序排列方案的优先次序，结果是：

$$e_{15}>e_{17}>e_{21}>e_{20}>e_{22}>e_{14}>e_{19}>e_{11}>e_{18}>e_5>e_{10}>e_6>$$
$$e_4>e_{12}>e_7>e_{16}>e_2>e_3>e_{13}>e_1>e_8$$

根据上面的计算结果：下标就是表示各个采样点序号，表示的结果就是在 Cd 这个主成分下，各个采样点由优到劣的排列顺序。

思考题 8

3. 答案：

第一步，对观测数据进行标准化处理，然后把标准化后的数据用矩阵 $\boldsymbol{X}$ 表示。

$$\boldsymbol{X}=\begin{pmatrix} -0.5355 & -1.2859 & 0.4346 & -0.8169 & 0.7853 \\ -0.7528 & -0.9285 & 2.7361 & -1.2593 & -0.6717 \\ -0.8615 & -0.3252 & -0.1799 & -0.7874 & -1.2644 \\ 2.3691 & -0.8126 & -0.4297 & 0.5397 & 0.6618 \\ -0.3576 & 0.3455 & -0.3763 & 0.0678 & -0.8693 \\ -0.4070 & 0.1107 & -0.3777 & 2.3681 & -1.2397 \\ -0.5058 & -0.4931 & -0.3833 & -0.3156 & 1.4768 \\ 0.5908 & 1.2493 & -0.4907 & -0.1091 & 1.0323 \\ 0.8575 & 1.9340 & -0.4854 & 0.3332 & -0.3754 \\ -0.3972 & 0.2058 & -0.4477 & -0.0206 & 0.4643 \end{pmatrix}$$

第二步，求样本的相关矩阵 $\boldsymbol{R}$。

$$\boldsymbol{R}=\begin{pmatrix} 1.0000 & 0.2376 & -0.3557 & 0.3048 & 0.3238 \\ 0.2376 & 1.0000 & -0.4798 & 0.3314 & -0.1133 \\ -0.3557 & -0.4798 & 1.0000 & -0.5321 & -0.2062 \\ 0.3048 & 0.3314 & -0.5321 & 1.0000 & -0.2390 \\ 0.3238 & -0.1133 & -0.2062 & -0.2390 & 1.0000 \end{pmatrix}$$

第三步，求 $\boldsymbol{R}$ 的特征值 λ 及其相应的特征向量。

$\boldsymbol{R}$ 的特征值及其累计方差贡献率分别为：

$\boldsymbol{R}$ 的特征值及其累计方差贡献率

特征值	2.145 9	1.314 7	0.700 9	0.587 2	0.251 2
累计方差贡献率/%	0.429 2	0.692 1	0.832 3	0.949 8	1.000 0

可以看出前三个特征根的累计方差贡献率已超过 75%，因此我们选择三个公共因子就可以了。

它们对应的特征向量：

$$\boldsymbol{e}_1=(0.4318\quad 0.4697\quad -0.5779\quad 0.5034\quad 0.0746)'$$

$$\boldsymbol{e}_2=(0.4115\quad -0.2304\quad -0.0710\quad -0.3396\quad 0.8107)'$$

$$\boldsymbol{e}_3=(-0.4017\quad 0.7101\quad -0.1614\quad -0.5285\quad 0.1702)'$$

第四步，求因子载荷矩阵 $\boldsymbol{A}$。

根据式(8.17)

$$A=(\sqrt{\lambda_1}e_1\ \sqrt{\lambda_2}e_2\ \sqrt{\lambda_3}e_3)=\begin{pmatrix}0.6325 & 0.4719 & -0.3363\\ 0.6881 & -0.2642 & 0.5945\\ -0.8466 & -0.0814 & -0.1351\\ 0.7375 & -0.3894 & -0.4425\\ 0.1093 & 0.9295 & 0.1425\end{pmatrix}$$

共同度，

$$h=\begin{pmatrix}0.7358\\ 0.8967\\ 0.7415\\ 0.8913\\ 0.8963\end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $g=(2.1459\quad 1.3147\quad 0.7009)$

第五步，对因子载荷矩阵 A 作正交旋转后得到的因子载荷矩阵为：

$$A_2^*=\begin{pmatrix}0.0901 & 0.5463 & -0.6552\\ 0.9377 & -0.0791 & -0.1059\\ -0.6469 & -0.2461 & 0.5124\\ 0.2485 & -0.2875 & -0.8642\\ -0.0023 & 0.9430 & 0.0842\end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $g=(1.3675\quad 1.3371\quad 1.4570)$

共同度不变，

$$h=\begin{pmatrix}0.7358\\ 0.8967\\ 0.7415\\ 0.8913\\ 0.8963\end{pmatrix}$$

第六步，求因子得分。

(1)求特殊向量方差 ψ。

$$\psi=\begin{pmatrix}0.2642 & 0.0000 & 0.0000 & 0.0000 & 0.0000\\ 0.0000 & 0.1033 & 0.0000 & 0.0000 & 0.0000\\ 0.0000 & 0.0000 & 0.2585 & 0.0000 & 0.0000\\ 0.0000 & 0.0000 & 0.0000 & 0.1087 & 0.0000\\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1037\end{pmatrix}$$

(2)运用式(8.30)，得到因子得分 F，结果见下表。

因子得分表

样本序号	因子得分		
	f_1	f_2	f_3
1	−0.990 1	0.618 0	0.600 8
2	−1.524 0	−0.750 5	1.263 1
3	0.052 1	−0.950 2	0.670 6
4	−1.108 7	1.292 1	−1.844 4
5	0.472 1	−0.739 0	0.066 3
6	−0.122 6	−1.665 0	−1.662 4
7	−0.046 5	1.058 2	0.485 8
8	1.206 0	0.985 4	0.278 6
9	1.598 9	−0.100 4	−0.153 1
10	0.462 9	0.251 4	0.294 6

4. 答案：

第一步，对观测数据进行标准化处理，然后把标准化后的数据用矩阵 $\boldsymbol{X}$ 表示。

$$\boldsymbol{X}=\begin{pmatrix} 0.0451 & -0.0332 & -0.9797 & -1.2907 & -0.7671 & -0.3670 \\ -0.3761 & -1.3163 & 1.2925 & -0.5575 & -0.5485 & -0.3977 \\ -1.0379 & 2.0022 & 0.6010 & 0.0535 & 0.0175 & -0.3231 \\ -1.2786 & 0.4535 & -0.0906 & -0.0484 & 2.2501 & -0.3523 \\ 1.7299 & -0.8296 & -1.0455 & 1.5809 & -0.7057 & -0.3272 \\ 0.5265 & -0.5641 & 1.2925 & 0.4608 & -0.4542 & 2.4741 \\ -0.4362 & 0.1880 & 0.0412 & 0.9700 & -0.2970 & -0.3377 \\ 0.8273 & 0.0996 & -1.1114 & -1.1685 & 0.5049 & -0.3691 \end{pmatrix}$$

第二步，求样本的相关矩阵 $\boldsymbol{R}$。

$$\boldsymbol{R}=\begin{pmatrix} 1.0000 & -0.5566 & -0.4434 & 0.2493 & -0.5195 & 0.2139 \\ -0.5566 & 1.0000 & -0.0673 & -0.0919 & 0.3770 & -0.2136 \\ -0.4434 & -0.0673 & 1.0000 & 0.1231 & -0.0819 & 0.5171 \\ 0.2493 & -0.0919 & 0.1231 & 1.0000 & -0.1450 & 0.2031 \\ -0.5195 & 0.3770 & -0.0819 & -0.1450 & 1.0000 & -0.1825 \\ 0.2139 & -0.2136 & 0.5171 & 0.2031 & -0.1825 & 1.0000 \end{pmatrix}$$

第三步，求 $\boldsymbol{R}$ 的特征值 λ 及其相应的特征向量。

R 的特征值及其累计方差贡献率分别为：

R 的特征值及其累计方差贡献率

特征值	2.182 8	1.606 7	0.919 5	0.646 8	0.568 9	0.075 2
累计方差贡献率/%	0.363 8	0.631 6	0.784 8	0.892 6	0.987 5	1.000 0

可以看出前三个特征根的累计方差贡献率已超过75%，因此我们选择三个公共因子就可以了。

它们对应的特征向量：

$$\boldsymbol{e}_1=(0.563\ 1\quad -0.504\ 4\quad 0.023\ 7\quad 0.277\ 1\quad -0.494\ 1\quad 0327\ 1)'$$

$$\boldsymbol{e}_2=(-0.355\ 6\quad 0.048\ 6\quad 0.743\ 0\quad 0.169\ 6\quad 0.032\ 0\quad 0.537\ 9)'$$

$$\boldsymbol{e}_3=(0.072\ 3\quad 0.343\ 6\quad -0.164\ 7\quad 0.899\ 6\quad 0.194\ 4\quad -0.051\ 0)'$$

第四步，求因子载荷矩阵 $\boldsymbol{A}$。

根据式(8.17)

$$\boldsymbol{A}=(\sqrt{\lambda_1}\boldsymbol{e}_1\ \sqrt{\lambda_2}\boldsymbol{e}_2\ \sqrt{\lambda_3}\boldsymbol{e}_3)=\begin{pmatrix} 0.831\ 9 & -0.450\ 8 & 0.069\ 3 \\ -0.745\ 3 & 0.061\ 6 & 0.329\ 5 \\ 0.035\ 0 & 0.941\ 8 & -0.157\ 9 \\ 0.409\ 4 & 0.215\ 0 & 0.862\ 6 \\ -0.730\ 1 & 0.040\ 6 & 0.186\ 4 \\ 0.483\ 3 & 0.681\ 8 & -0.048\ 9 \end{pmatrix}$$

共同度，
$$\boldsymbol{h}=\begin{pmatrix} 0.900\ 0 \\ 0.667\ 8 \\ 0.913\ 2 \\ 0.957\ 9 \\ 0.569\ 4 \\ 0.700\ 8 \end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(2.182\ 8\quad 1.606\ 7\quad 0.919\ 5)$

第五步，对因子载荷矩阵 $\boldsymbol{A}$ 作正交旋转后得到的因子载荷矩阵为：

$$\boldsymbol{A}_2^*=\begin{pmatrix} 0.856\ 5 & -0.293\ 9 & 0.283\ 1 \\ -0.801\ 6 & -0.142\ 3 & 0.070\ 0 \\ -0.144\ 2 & 0.944\ 5 & -0.018\ 7 \\ 0.073\ 7 & 0.115\ 7 & 0.969\ 1 \\ -0.740\ 4 & -0.132\ 0 & -0.061\ 4 \\ 0.300\ 2 & 0.755\ 7 & 0.198\ 8 \end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(2.040\ 6\quad 1.600\ 6\quad 1.067\ 8)$

第六步，求因子得分。

(1)求特殊向量方差 $\boldsymbol{\psi}$。

$$\boldsymbol{\psi}=\begin{pmatrix} 0.1000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3322 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0868 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0421 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4306 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.2992 \end{pmatrix}$$

(2)运用式(8.30)，得到因子得分 $\boldsymbol{F}$，结果见下表。

因子得分表

样本序号	因子得分		
	f_1	f_2	f_3
1	−0.021 7	0.007 4	0.064 3
2	−0.033 8	0.044 0	0.119 8
3	−0.086 2	0.039 1	0.199 8
4	−0.124 8	0.013 6	0.194 0
5	−0.039 9	0.005 2	0.336 3
6	0.113 9	0.668 5	0.326 9
7	−0.068 8	0.027 4	0.277 4
8	−0.049 7	−0.005 0	0.085 4

5. 答案：

第一步，对观测数据进行标准化处理，然后把标准化后的数据用矩阵 $\boldsymbol{X}$ 表示。

$$\boldsymbol{X}=\begin{pmatrix} 1.1682 & 1.7279 & -0.5658 & -0.5287 & -0.5866 & 2.3020 & -0.4214 \\ 1.3811 & 1.3663 & -0.7435 & -0.9154 & -1.0987 & -0.1703 & -0.7568 \\ -0.5655 & -0.5754 & -0.5777 & 0.2009 & 0.1282 & -0.6792 & -0.5902 \\ 0.0023 & 0.0501 & -0.7435 & -0.5830 & -0.6020 & -0.5285 & -0.7547 \\ -0.9913 & -0.8837 & 1.6852 & 1.6497 & 1.6952 & 0.5825 & 1.5933 \\ -0.9812 & -0.8867 & 0.7273 & 0.8511 & 0.7292 & -0.5012 & 0.6415 \\ -1.0522 & -0.9756 & 1.4517 & 1.1668 & 1.2352 & -0.5166 & 1.5871 \\ 1.1581 & 0.4621 & -0.5753 & -0.9196 & -0.8316 & 0.4465 & -0.5514 \\ -0.1194 & -0.2849 & -0.6584 & -0.9217 & -0.6688 & -0.9351 & -0.7474 \end{pmatrix}$$

第二步，求样本的相关矩阵 $\boldsymbol{R}$。

$$\boldsymbol{R}=\begin{pmatrix} 1.0000 & 0.9466 & -0.7328 & -0.8312 & -0.8569 & 0.4904 & -0.7042 \\ 0.9466 & 1.0000 & -0.6659 & -0.7278 & -0.7703 & 0.6270 & -0.6285 \\ -0.7328 & -0.6659 & 1.0000 & 0.9407 & 0.9526 & -0.0206 & 0.9960 \\ -0.8312 & -0.7278 & 0.9407 & 1.0000 & 0.9915 & -0.0633 & 0.9324 \\ -0.8569 & -0.7703 & 0.9526 & 0.9915 & 1.0000 & -0.0868 & 0.9431 \\ 0.4904 & 0.6270 & -0.0206 & -0.0633 & -0.0868 & 1.0000 & 0.0247 \\ -0.7042 & -0.6285 & 0.9960 & 0.9324 & 0.9431 & 0.0247 & 1.0000 \end{pmatrix}$$

第三步，求 $\boldsymbol{R}$ 的特征值 λ 及其相应的特征向量。

$\boldsymbol{R}$ 的特征值及其累计方差贡献率分别为：

$\boldsymbol{R}$ 的特征值及其累计方差贡献率

特征值	5.276 9	1.446 3	0.187 0	0.058 5	0.026 5	0.002 6	0.002 1
累计方差贡献率/%	0.753 8	0.960 5	0.987 2	0.995 5	0.999 3	0.999 7	1.000 0

可以看出前两个特征根的累计方差贡献率已超过85%，因此我们选择两个公共因子就可以了。

它们对应的特征向量：

$\boldsymbol{e}_1=(-0.3995\quad -0.3761\quad 0.4083\quad 0.4195\quad 0.4268\quad -0.1155\quad 0.4010)'$

$\boldsymbol{e}_2=(0.2601\quad 0.3841\quad 0.2363\quad 0.1682\quad 0.1435\quad 0.7778\quad 0.2740)'$

第四步，求因子载荷矩阵 $\boldsymbol{A}$。

根据式(8.17)

$$\boldsymbol{A}=(\sqrt{\lambda_1}\boldsymbol{e}_1\ \ \sqrt{\lambda_2}\boldsymbol{e}_2)=\begin{pmatrix} -0.9177 & 0.3129 \\ -0.8639 & 0.4619 \\ 0.9379 & 0.2842 \\ 0.9637 & 0.2023 \\ 0.9803 & 0.1726 \\ -0.2652 & 0.9355 \\ 0.9212 & 0.3295 \end{pmatrix}$$

共同度，

$$\boldsymbol{h}=\begin{pmatrix} 0.9401 \\ 0.9597 \\ 0.9604 \\ 0.9696 \\ 0.9908 \\ 0.9454 \\ 0.9572 \end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(5.2769\quad 1.4463)$

第五步，对因子载荷矩阵 $\boldsymbol{A}$ 作正交旋转后得到的因子载荷矩阵为：

$$\boldsymbol{A}_2^*=\begin{pmatrix} -0.7354 & 0.6319 \\ -0.6300 & 0.7502 \\ 0.9763 & -0.0852 \\ 0.9698 & -0.1708 \\ 0.9742 & -0.2045 \\ 0.1019 & 0.9670 \\ 0.9777 & -0.0370 \end{pmatrix}$$

各个公因子 f_j 对所有变量的贡献 $\boldsymbol{g}=(4.746\,5 \quad 1.976\,7)$

第六步，求因子得分。

(1)求特殊向量方差 $\boldsymbol{\psi}$。

$$\boldsymbol{\psi}=\begin{pmatrix} 0.059\,9 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 \\ 0.000\,0 & 0.040\,3 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 \\ 0.000\,0 & 0.000\,0 & 0.039\,6 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 \\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.030\,4 & 0.000\,0 & 0.000\,0 & 0.000\,0 \\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.009\,2 & 0.000\,0 & 0.000\,0 \\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.054\,6 & 0.000\,0 \\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.042\,8 \end{pmatrix}$$

(2)运用式(8.30)，得到因子得分 $\boldsymbol{F}$，结果见下表。

因子得分表

样本序号	因子得分	
	f_1	f_2
1	1.208 6	0.839 4
2	0.755 4	0.545 3
3	1.658 0	0.678 1
4	1.114 3	0.584 6
5	2.894 9	1.033 8
6	2.131 3	0.791 2
7	2.519 4	0.869 4
8	0.971 8	0.639 0
9	1.051 0	0.538 6
10	1.208 6	0.839 4

思考题 10

1. 答案：当一个变量呈现为空间分布时，就称之为区域化变量(regionalized variable)，区域化随机变量与普通随机变量不同，普通随机变量的取值符合某种概率分布，而区域化随机变量则根据其在一个域内的位置不同而取值。也就是说，区域化随机变量是变通随机变量在一个域内确定位置上的特定取值，它是与位置有关的随机函数。

区域化随机变量之间的差异，可以用空间协方差来表示。协方差又叫做半方差，是地统计学中的关键概念。在概率论中，随机向量(x, y)的协方差被定义为：

$$\mathrm{Cov}(x, y)=E[(x-E(x))(y-E(y))]$$

将环境空间信息看作成随空间位置 x 而变化的区域化变量 $Z(x)$(为讨论问题方便不妨设 $Z(x)$定义在一维坐标轴上)，那么，当空间点 x 在一维 x 轴上变化时，区域化变量 $Z(x)$在点

x 和 $x+h$ 处的值 $Z(x)$ 与 $Z(x+h)$ 差的方差一半定义为区域化变量 $Z(x)$ 在 x 轴方向上的变差函数，记作 $\gamma(x, h)$。即：

$$\gamma(x,\ h)=\frac{1}{2}D[Z(x)-Z(x+h)]$$

根据协方差函数的理论，变差函数可以展开为：

$$\begin{aligned}\gamma(x,\ h)&=\frac{1}{2}D[Z(x)-Z(x+h)]\\&=\frac{1}{2}E[Z(x)-Z(x+h)]^2-\frac{1}{2}\{E[Z(x)]-E[Z(x+h)]\}^2\end{aligned}$$

2. 答案：克立格法(Kriging)也称空间局部估计或空间局部插值，是空间统计学中两大主要方法之一。它是建立在变差函数理论及结构分析基础上的，在有限区域内对区域化变量的取值进行无偏最优估计的一种方法。这种方法最早由南非矿业工程师克立格和统计学家西舍尔在 20 世纪 50 年代根据样本空间位置不同和样本的相关程度的不同，对每个样本赋予一定的权重，进行滑动加权平均，来估计未知样点上样本平均值的一种方法。

克立格法实质上是利用区域化变量的原始数据和变差函数的结构特点，对未采样点的区域化变量的取值进行线性无偏最优估计的一种方法。从数学的角度讲就是一种对空间分布的数据求线性最优无偏内插估计量(best linear unbiased estimator，简写为 BLUE)的一种方法。更具体地讲，它是根据待估样点(或待估块段)有限邻域内若干已测定的样点数据，在认真考虑了样点的形状、大小和空间相互位置关系，它们与待估样点间相互空间位置关系以及变差函数提供的结构信息之后，对该待估样点值进行的一种线性无偏最优估计。

在环境科学中，经常遇到通过采样数据推求污染物空间分布的现象，比如土壤污染监测，这时可以根据已测定的样点数据，通过克立格法求线性最优无偏内插估计量，估算污染物在整个区域的空间特征。

3. 答案：

$$\begin{aligned}\gamma(1)=\frac{1}{2\times36}[&(15-18)^2+(18-16)^2+(16-15)^2+(15-10)^2+(13-15)^2+\\&(15-20)^2+(20-17)^2+(17-16)^2+(10-11)^2+(18-21)^2+(12-14)^2+\\&(14-15)^2+(15-18)^2+(18-16)^2+(17-19)^2+(19-23)^2+(23-21)^2+\\&(21-18)^2+(15-13)^2+(13-10)^2+(10-12)^2+(12-17)^2+(18-15)^2+\\&(15-11)^2+(11-14)^2+(14-19)^2+(16-20)^2+(15-23)^2+(15-17)^2+\\&(17-18)^2+(18-18)^2+(18-21)^2+(10-16)^2+(16-21)^2+(21-16)^2+\\&(16-18)^2]=\frac{424}{72}=5.89\end{aligned}$$

$$\begin{aligned}\gamma(2)=\frac{1}{2\times26}[&(15-16)^2+(18-15)^2+(16-10)^2+(13-20)^2+(15-17)^2+\\&(20-16)^2+(11-18)^2+(12-15)^2+(14-18)^2+(15-16)^2+(17-23)^2+\\&(19-21)^2+(23-18)^2+(15-10)^2+(13-12)^2+(10-17)^2+(18-11)^2+\\&(15-14)^2+(11-19)^2+(20-15)^2+(15-18)^2+(17-18)^2+(18-21)^2+\\&(10-21)^2+(16-16)^2+(21-18)^2]=\frac{618}{52}=11.88\end{aligned}$$

$$\gamma(3)=\frac{1}{2\times 20}[(15-15)^2+(18-10)^2+(13-17)^2+(15-16)^2+(10-18)^2+$$
$$(11-21)^2+(12-18)^2+(14-16)^2+(17-21)^2+(19-18)^2+(15-12)^2+$$
$$(13-17)^2+(18-14)^2+(15-19)^2+(16-15)^2+(20-23)^2+(15-18)^2+$$
$$(17-21)^2+(10-16)^2+(16-18)^2]=\frac{434}{40}=10.85$$

$$\gamma(4)=\frac{1}{2\times 10}[(15-10)^2+(13-16)^2+(10-21)^2+(12-16)^2+(17-18)^2+$$
$$(15-17)^2+(18-19)^2+(16-23)^2+(15-21)^2+(10-18)^2]$$
$$=\frac{326}{20}=16.30$$

4. 答案：

根据公式(10.32)，可以计算：

$$\bar{y}=\frac{36\times 5.89+26\times 11.88+20\times 10.85+10\times 16.30}{36+26+20+10}=9.79$$

$$\overline{x_1}=\frac{36\times 2+26\times 4+20\times 6+10\times 8}{36+26+20+10}=4.09$$

$$\overline{x_2}=\frac{36\times 2^3+26\times 4^3+20\times 6^3+10\times 8^3}{36+26+20+10}=123.83$$

$$L_{11}=36\times(2-4.09)^2+26\times(4-4.09)^2+20\times(6-4.09)^2+$$
$$10\times(8-4.09)^2=383.30$$

$$L_{22}=36\times(8-123.83)^2+26\times(64-123.83)^2+20\times(216-123.83)^2+$$
$$10\times(512-123.83)^2=2\ 252\ 733.22$$

$$L_{12}=L_{21}=36\times(2-4.09)\times(8-123.83)+26\times(4-4.09)\times(64-123.83)+20\times(6-4.09)\times$$
$$(216-123.83)+10\times(8-4.09)\times(512-123.83)=27\ 553.39$$

$$L_{1y}=36\times(2-4.09)\times(5.89-9.79)+26\times(4-4.09)\times(11.88-9.79)+20\times(6-4.09)\times$$
$$(10.85-9.79)+10\times(8-4.09)\times(16.30-9.79)=583.56$$

$$L_{2y}=36\times(8-123.83)\times(5.89-9.79)+26\times(64-123.83)\times(11.88-9.79)+20\times(216-$$
$$123.83)\times(10.85-9.79)+10\times(512-123.83)\times(16.30-9.79)=40\ 235.24$$

$$L_{yy}=36\times(5.89-9.79)^2+26\times(11.88-9.79)^2+20\times(10.85-9.79)^2+10\times(16.30-9.79)^2$$
$$=1\ 107.40$$

从而可求得，

$$b_1=\frac{583.56\times 2\ 252\ 733.22-40\ 235.24\times 27\ 553.39}{383.30\times 2\ 252\ 733.22-27\ 553.39\times 27\ 553.39}=1.98$$

$$b_2=\frac{40\ 235.24\times 383.30-583.56\times 27\ 553.39}{383.30\times 2\ 252\ 733.22-27\ 553.39\times 27\ 553.39}=-0.01$$

$$b_0=9.79-1.98\times 4.09+0.01\times 123.83=2.93$$

由于 $b_0>0$，$b_1>0$，$b_2<0$，此时球状模型中三个参数 C_0，C 和 a 分别为：

$$C_0=b_0=2.93$$

$$a=\sqrt{\frac{-1.98}{3\times(-0.01)}}=8.41$$

$$C=\frac{2\times1.98}{3}\sqrt{\frac{-1.98}{3\times(-0.01)}}=11.10$$

因此，球状模型为：

$$\gamma(h)=\begin{cases}0 & (h=0)\\ 2.93+11.10\times\left(\frac{3}{2}\times\frac{h}{8.41}-\frac{1}{2}\times\frac{h^3}{8.41^3}\right) & (0<h<8.41)\\ 14.03 & (h>8.41)\end{cases}$$

5. 答案：

根据协方差与变差函数之间的关系以及第4题求得的变差函数，可得协方差函数：

$$c^*(h)=\begin{cases}14.03 & (h=0)\\ 11.10\times\left[1-\left(\frac{3}{2}\times\frac{h}{8.41}-\frac{1}{2}\times\frac{h^3}{8.41^3}\right)\right] & (0<h<8.41)\\ 0 & (h>8.41)\end{cases}$$

因此，

$$c_{11}=c_{22}=c_{33}=c_{44}=14.03$$

$$c_{12}=c_{21}=c_{13}=c_{31}=c_{04}=14.03-\gamma(\sqrt{4^2+2^2})$$

$$=14.03-\gamma(2\sqrt{5})=14.03-\left[2.93+11.10\times\left(\frac{3}{2}\times\frac{2\sqrt{5}}{8.41}-\frac{1}{2}\times\frac{(2\sqrt{5})^3}{8.41^3}\right)\right]$$

$$=3.11$$

$$c_{14}=c_{41}=c_{02}=14.03-\gamma(\sqrt{4^2})=14.03-\gamma(4)$$

$$=14.03-\left[2.93+11.10\times\left(\frac{3}{2}\times\frac{4}{8.41}-\frac{1}{2}\times\frac{4^3}{8.41^3}\right)\right]=3.77$$

$$c_{01}=14.03-\gamma(\sqrt{2^2})=14.03-\gamma(2)$$

$$=14.03-\left[2.93+11.10\times\left(\frac{3}{2}\times\frac{2}{8.41}-\frac{1}{2}\times\frac{2^3}{8.41^3}\right)\right]=7.21$$

$$c_{23}=c_{32}=c_{03}=14.03-\gamma(\sqrt{2^2+2^2})=14.03-\gamma(2\sqrt{2})$$

$$=14.03-\left[2.93+11.10\times\left(\frac{3}{2}\times\frac{2\sqrt{2}}{8.41}-\frac{1}{2}\times\frac{(2\sqrt{2})^3}{8.41^3}\right)\right]=5.66$$

$$c_{24}=c_{42}=14.03-\gamma(\sqrt{8^2+2^2})=14.03-\gamma(2\sqrt{17})$$

$$=14.03-\left[2.93+11.10\times\left(\frac{3}{2}\times\frac{2\sqrt{17}}{8.41}-\frac{1}{2}\times\frac{(2\sqrt{17})^3}{8.41^3}\right)\right]=0.00$$

$$c_{34}=c_{43}=14.03-\gamma(\sqrt{6^2+4^2})=14.03-\gamma(2\sqrt{13})$$

$$=14.03-\left[2.93+11.10\times\left(\frac{3}{2}\times\frac{2\sqrt{13}}{8.41}-\frac{1}{2}\times\frac{(2\sqrt{13})^3}{8.41^3}\right)\right]=0.33$$

可以得到克立格方程组：

$$\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\\\lambda_4\\\mu\end{pmatrix}=\begin{pmatrix}14.03 & 3.11 & 3.11 & 3.77 & 1.00\\3.11 & 14.03 & 5.66 & 0.00 & 1.00\\3.11 & 5.66 & 14.03 & 0.33 & 1.00\\3.77 & 0.00 & 0.33 & 14.03 & 1.00\\1.00 & 1.00 & 1.00 & 1.00 & 0.00\end{pmatrix}^{-1}\begin{pmatrix}7.21\\3.77\\5.66\\3.11\\1.00\end{pmatrix}=\begin{pmatrix}0.40\\0.06\\0.28\\0.09\\0.17\end{pmatrix}$$

所以，x_0 点的克立格估计值为：

$$Z_0^*=20\times0.40-21\times0.06+18\times0.28+13\times0.09=15.47$$

克立格估计方差为：

$$\sigma_K^2=14.03-(0.40\times7.21-0.06\times3.77+0.28\times5.66+0.09\times3.11)-0.17=8.89$$

6. 答案：主成分分析就是设法将原来众多的具有一定相关性的指标，重新组合成一组新的相互无关的综合指标来代替原来的指标，而保持其原指标所提供的大量信息。

环境空间主成分分析法则是在空间数据的基础上，通过将原始空间坐标轴旋转，将相关的多变量环境空间数据转化为少数几个不相关的综合指标，实现用较少的综合指标最大限度地保留原来较多环境变量所反映的信息。空间主成分分析是在地理信息系统软件 ARC/INFO 的 GRID 模块支持下，利用该模块中的 PRINCOMP 函数，通过对原始空间轴的旋转完成主成分分析。在提取出来的空间主成分的基础上，可以进行其他方面的工作，比如区域生态环境综合评价、区域生态脆弱性评价等。

7. 答案：环境空间主成分分析的重要步骤为：

(1)在 ARC/INFO 中用 POLYGRID 命令将环境矢量数据转化为栅格数据。

(2)按照一定的标准化方法对转化生成的栅格数据进行标准化处理。

(3)利用 GRID 模块中的 MAKESTACK 命令将标准化处理后的指标 X_i 图转化为一个综合图。

(4)利用 GRID 模块中的 PRINCOMP 函数，对综合图进行主成分转换，根据所转换的空间主成分特征向量，利用公式：

$$a_i=\lambda_i/\sum_{i=1}^{m}\lambda_i$$

计算得到各主成分的贡献率，再根据主成分累计贡献率大小，来确定主成分数。

(5)在环境综合评价中，综合评价指数定义为 M 个主成分的加权和，而权重用每个主成分相对应的贡献率来表示，即：

$$E=a_1Y_1+a_2Y_2+\cdots+a_jY_j\quad(j=1,\ 2,\ \cdots,\ M)$$

其中：E 为环境综合评价指数；Y_j 为第 j 个主成分；a_j 为第 j 主成分对应的贡献率。

8. 答案：实现用较少的综合指标最大限度地保留原来较多环境变量所反映的信息。

9. 答案：

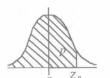
(1)用 POLYGRID 命令将环境矢量数据转化为栅格数据。

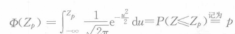
(2)利用 GRID 模块中的 MAKESTACK 命令将指标图转化为一个综合图。

(3)利用 GRID 模块中的 PRINCOMP 函数，对综合图进行主成分转换。

附　录

附表 1　标准正态分布表





$$\Phi(Z_p)=\int_{-\infty}^{Z_p}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du=P(Z\leqslant Z_p)\overset{记为}{=}p$$

$p=1-\alpha$	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.000 000	0.002 507	0.005 013	0.007 520	0.010 027	0.012 533	0.015 040	0.017 547	0.020 054	0.022 562
0.51	0.025 069	0.027 576	0.030 084	0.032 592	0.035 100	0.037 608	0.040 117	0.042 626	0.045 135	0.047 644
0.52	0.050 154	0.052 664	0.055 174	0.057 684	0.060 195	0.062 707	0.065 219	0.067 731	0.070 243	0.072 756
0.53	0.075 270	0.077 784	0.080 298	0.082 813	0.085 329	0.087 845	0.090 361	0.092 879	0.095 396	0.097 915
0.54	0.100 434	0.102 953	0.105 474	0.107 995	0.110 516	0.113 039	0.115 562	0.118 085	0.120 610	0.123 135
0.55	0.125 661	0.128 188	0.130 716	0.133 245	0.135 774	0.138 304	0.140 835	0.143 367	0.145 900	0.148 434
0.56	0.150 969	0.153 505	0.156 042	0.158 580	0.161 119	0.163 658	0.166 199	0.168 741	0.171 285	0.173 829
0.57	0.176 374	0.178 921	0.181 468	0.184 017	0.186 567	0.189 118	0.191 671	0.194 225	0.196 780	0.199 336
0.58	0.201 893	0.204 452	0.207 013	0.209 574	0.212 137	0.214 702	0.217 267	0.219 835	0.222 403	0.224 973
0.59	0.227 545	0.230 118	0.232 693	0.235 269	0.237 847	0.240 426	0.243 007	0.245 590	0.248 174	0.250 760
0.60	0.253 347	0.255 936	0.258 527	0.261 120	0.263 714	0.266 311	0.268 909	0.271 508	0.274 110	0.276 714
0.61	0.279 319	0.281 926	0.284 536	0.287 147	0.289 760	0.292 375	0.294 992	0.297 611	0.300 232	0.302 855
0.62	0.305 481	0.308 108	0.310 738	0.313 369	0.316 003	0.318 639	0.321 278	0.323 918	0.326 561	0.329 206
0.63	0.331 853	0.334 503	0.337 155	0.339 809	0.342 766	0.345 126	0.347 787	0.350 451	0.353 118	0.355 787
0.64	0.358 459	0.361 133	0.363 810	0.366 489	0.369 171	0.371 856	0.374 543	0.377 234	0.379 926	0.382 622
0.65	0.385 320	0.388 022	0.390 786	0.393 433	0.396 142	0.398 855	0.401 571	0.404 289	0.407 011	0.409 735
0.66	0.412 463	0.415 194	0.417 928	0.420 665	0.423 405	0.426 148	0.428 895	0.431 644	0.434 397	0.437 154
0.67	0.439 913	0.442 676	0.445 443	0.448 212	0.450 985	0.453 762	0.456 542	0.459 326	0.462 113	0.464 904
0.68	0.467 699	0.470 497	0.473 299	0.476 104	0.478 914	0.481 727	0.484 544	0.487 365	0.490 189	0.493 018
0.69	0.495 850	0.498 687	0.501 527	0.504 372	0.507 221	0.510 073	0.512 930	0.515 792	0.518 657	0.521 527
0.70	0.524 401	0.527 279	0.530 161	0.533 049	0.535 940	0.538 836	0.541 737	0.544 642	0.547 551	0.550 466
0.71	0.553 385	0.556 308	0.559 237	0.562 175	0.655 108	0.568 051	0.570 999	0.573 952	0.576 910	0.579 873
0.72	0.582 842	0.585 815	0.588 793	0.591 777	0.594 766	0.597 760	0.600 760	0.603 765	0.606 775	0.609 792
0.73	0.612 813	0.615 840	0.618 873	0.621 912	0.624 956	0.628 006	0.631 062	0.634 124	0.637 192	0.640 266
0.74	0.643 345	0.646 431	0.649 524	0.652 622	0.655 727	0.658 838	0.661 955	0.665 079	0.668 209	0.671 346
0.75	0.674 490	0.677 640	0.680 797	0.683 961	0.687 131	0.690 309	0.693 493	0.696 685	0.699 884	0.703 089
0.76	0.706 303	0.709 523	0.712 751	0.715 986	0.719 229	0.722 479	0.725 737	0.729 003	0.732 276	0.735 558
0.77	0.738 847	0.742 144	0.745 450	0.748 763	0.752 085	0.755 415	0.758 754	0.762 101	0.765 456	0.768 820
0.78	0.772 193	0.775 575	0.778 966	0.782 365	0.785 774	0.789 192	0.792 619	0.796 055	0.799 501	0.802 956
0.79	0.806 421	0.809 896	0.813 380	0.816 875	0.820 379	0.823 894	0.827 418	0.830 953	0.834 499	0.838 055
0.80	0.841 621	0.845 199	0.848 787	0.852 386	0.855 996	0.859 617	0.863 250	0.866 894	0.879 550	0.874 217
0.81	0.877 896	0.881 587	0.885 290	0.889 006	0.892 733	0.896 473	0.900 226	0.903 991	0.907 770	0.911 561
0.82	0.915 365	0.919 183	0.923 014	0.926 859	0.930 717	0.934 589	0.938 476	0.942 376	0.946 291	0.950 221
0.83	0.954 165	0.958 124	0.962 099	0.966 088	0.970 093	0.974 114	0.978 150	0.982 203	0.986 271	0.990 356
0.84	0.994 458	0.998 576	1.002 712	1.006 864	1.011 034	1.015 222	1.019 428	1.023 651	1.027 893	1.032 154
0.85	1.036 433	1.040 732	1.045 050	1.049 387	1.053 744	1.058 122	1.062 519	1.066 938	1.071 377	1.075 837
0.86	1.080 319	1.084 823	1.089 349	1.097 897	1.098 468	1.103 063	1.107 680	1.112 321	1.116 987	1.121 677

续表

$p=1-\alpha$	0. 000	0. 001	0. 002	0. 003	0. 004	0. 005	0. 006	0. 007	0. 008	0. 009
0. 87	1. 126 391	1. 131 131	1. 135 896	1. 140 687	1. 145 505	1. 150 349	1. 155 221	1. 160 120	1. 165 047	1. 170 002
0. 88	1. 174 987	1. 180 001	1. 185 044	1. 190 118	1. 195 223	1. 200 359	1. 205 527	1. 210 727	1. 215 960	1. 221 227
0. 89	1. 226 528	1. 231 864	1. 237 235	1. 242 641	1. 248 085	1. 253 565	1. 259 084	1. 264 641	1. 270 238	1. 275 874
0. 90	1. 281 552	1. 287 271	1. 293 032	1. 298 837	1. 304 685	1. 310 579	1. 316 519	1. 322 505	1. 328 539	1. 334 622
0. 91	1. 340 755	1. 346 939	1. 353 174	1. 359 463	1. 365 806	1. 372 204	1. 378 659	1. 385 172	1. 391 744	1. 398 377
0. 92	1. 405 072	1. 411 830	1. 418 654	1. 425 544	1. 432 503	1. 439 531	1. 446 632	1. 453 806	1. 461 056	1. 468 384
0. 93	1. 475 791	1. 483 280	1. 490 853	1. 498 513	1. 506 262	1. 514 102	1. 522 036	1. 530 068	1. 538 199	1. 546 433
0. 94	1. 554 774	1. 563 224	1. 571 787	1. 580 467	1. 589 268	1. 598 193	1. 607 248	1. 616 436	1. 625 763	1. 635 234
0. 95	1. 644 854	1. 654 628	1. 664 563	1. 674 665	1. 684 941	1. 695 398	1. 706 043	1. 716 886	1. 727 934	1. 739 198
0. 96	1. 750 686	1. 762 410	1. 774 382	1. 786 613	1. 799 118	1. 811 911	1. 825 007	1. 838 424	1. 852 180	1. 866 296
0. 97	1. 880 794	1. 895 698	1. 911 036	1. 926 837	1. 943 134	1. 959 964	1. 977 368	1. 995 393	2. 014 091	2. 033 520
0. 98	2. 053 749	2. 074 855	2. 096 927	2. 120 072	2. 144 411	2. 170 090	2. 197 286	2. 226 212	2. 257 129	2. 290 368
0. 99	2. 326 348	2. 365 618	2. 408 916	2. 457 263	2. 512 144	2. 575 829	2. 652 070	2. 747 781	2. 878 162	3. 090 232

注：本表对于下侧概率给出正态分布的分位数 Z_p。

例：对于 $p=0.95, Z_p=1.644\ 854$。

附表 2　相关系数检验表

$n-2$	5%	1%	$n-2$	5%	1%	$n-2$	5%	1%
1	0. 997	1. 000	16	0. 468	0. 590	35	0. 325	0. 418
2	0. 950	0. 990	17	0. 456	0. 575	40	0. 304	0. 393
3	0. 878	0. 959	18	0. 444	0. 561	45	0. 288	0. 372
4	0. 811	0. 917	19	0. 433	0. 549	50	0. 273	0. 354
5	0. 754	0. 874	20	0. 423	0. 537	60	0. 250	0. 325
6	0. 707	0. 834	21	0. 413	0. 526	70	0. 232	0. 302
7	0. 666	0. 798	22	0. 404	0. 515	80	0. 217	0. 283
8	0. 632	0. 765	23	9. 396	0. 505	90	0. 205	0. 267
9	0. 602	0. 735	24	0. 388	0. 496	100	0. 195	0. 254
10	0. 576	0. 708	25	0. 381	0. 487	125	0. 174	0. 228
11	0. 553	0. 684	26	0. 374	0. 478	150	0. 159	0. 208
12	0. 532	0. 661	27	0. 367	0. 470	200	0. 138	0. 181
13	0. 514	0. 641	28	0. 361	0. 463	300	0. 113	0. 148
14	0. 497	0. 623	29	0. 355	0. 456	400	0. 098	0. 128
15	0. 482	0. 606	30	0. 349	0. 449	1 000	0. 062	0. 081

附表3　χ^2分布临界值表

例如：自由度 $n=20$，$P(\chi^2>34.17)=0.025$。

n	$\alpha=0.995$	$\alpha=0.990$	$\alpha=0.975$	$\alpha=0.950$	$\alpha=0.050$	$\alpha=0.025$	$\alpha=0.010$	$\alpha=0.005$
1	0.000 039 3	0.000 157	0.000 982	0.003 93	3.841	5.024	6.635	7.879
2	0.010	0.020 1	0.050 6	0.103	5.991	7.378	9.210	10.579
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.186	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672

附表 4　t 分布临界值表

例如：自由度 $n=20$，$P(t>1.725)=0.05$。

n \ α	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.000 5
1	0.100	1.376	1.963	3.076	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.397	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.733	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

附表 5 *F* 分布临界值表

例如：自由度 $n_1=10$，$n_2=29$，$P(F>2.18)=0.05$，$P(F>3.00)=0.01$。

注：n_2 下面的数字是 1%的显著性水平，n_2 上面的数字是 5%的显著性水平。

分母的自由度 n_2 \ 分子的自由度 n_1	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
	4 052	4 999	5 403	5 625	5 764	5 859	5 928	5 981	6 022	6 056	6 082	6 106
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	18.01	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.23	2.25
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12

续表

n_2 \ n_1	分子的自由度											
分母的自由度	1	2	3	4	5	6	7	8	9	10	11	12
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.32	2.28	2.24	2.20	2.16
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.34	3.21	3.13	3.05	2.99
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80
34	4.13	3.28	2.88	2.65	2.49	2.38	3.30	2.23	2.17	2.12	2.08	2.50
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03
	7.39	5.25	4.38	3.80	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64
44	4.06	3.21	2.82	2.58	2.43	2.34	2.23	2.16	2.10	2.05	2.01	1.98
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93
	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90
	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83
	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.65	2.56	2.47	2.40	2.33
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82
	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30

续表

	n_2 \ n_1	分子的自由度											
		1	2	3	4	5	6	7	8	9	10	11	12
分母的自由度	200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80
		6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28
	400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78
		6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.33
	1 000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76
		6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20
	∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75
		6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18

	n_2 \ n_1	分子的自由度											
		14	16	20	24	30	40	50	75	100	200	500	∞
分母的自由度	1	245	246	248	249	250	251	252	253	253	254	254	254
		6 142	6 169	6 208	6 234	6 258	6 286	6 302	6 323	6·334	6 352	6 361	6 366
	2	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
		99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50
	3	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
		26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12
	4	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
		14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46
	5	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
		9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02
	6	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
		7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88
	7	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
		6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65
	8	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
		5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86
	9	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
		5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31
	10	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
		4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.94
	11	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
		4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60
	12	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
		4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36
	13	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
		3.85	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16
	14	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
		3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
	15	2.48	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
		3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
	16	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
		3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.89	2.86	2.80	2.77	2.75
	17	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
		3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65
	18	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
		3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57
	19	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
		3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49

续表

分母的自由度 n_2 \ n_1	分子的自由度											
	14	16	20	24	30	40	50	75	100	200	500	∞
20	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
21	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
22	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
23	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	2.97	2.89	2.78	2.79	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
24	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21
25	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
26	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13
27	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10
28	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06
29	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03
30	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
32	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96
34	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91
36	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	2.62	3.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87
38	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84
40	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81
42	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
48	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
50	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
55	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	2.43	2.35	2.23	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
60	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
65	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	2.37	2.30	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
70	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49

续表

	n_1 / n_2	分子的自由度											
		14	16	20	24	30	40	50	75	100	200	500	∞
分母的自由度	100	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
		2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
	125	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
		2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37
	150	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
		2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33
	200	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
		2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
	400	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
		2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19
	1 000	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
		2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
	∞	1.67	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
		2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00